VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

ELPOD: Podnikání na internetu, směry, možnosti, řešení, aplikace

ELPOD: Colecting Information from Ecommerce Sources for Future Distribution

Ostrava 2010 Bc. Tomáš Sobotík

Zadání …

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 30. dubna 2010

……
Tomáš Sobotík

Poděkování

Děkuji tímto vedoucímu diplomové práce Ing. Radoslavu Fasugovi, Ph.D. za odbornou pomoc a konzultaci při realizaci tohoto projektu.

Abstrakt a klíčová slova

Abstrakt:

Tato práce se zabývá problematikou elektronické komerce a to z několika pohledů. Tematické zaměření je podnikání na internetu, jeho možnosti a rozšíření. Práce obsahuje přehled nejčastějších forem elektronické komerce, včetně právních náležitostí související s problematikou. Druhá část textu se poté věnuje pokročilejším formám a možnostem, jak podnikání zefektivnit a dále rozvíjet. Jde zejména o problematiku datových skladů a data miningu v návaznosti na internetové podnikání. Práce tedy obsahuje popis procesu budování datového skladu v prostředí Microsoft SQL Server 2005 – jednotlivé dílčí fáze až po závěrečné reporty a využití nově získaných informací. Dále popis nejpoužívanějších data miningových metod, jejich praktickou aplikaci v prostředí elektronické komerce s využitím jak komerčních nástrojů (Microsoft Business Intelligence Development Studio), tak open source nástrojů (RapidMiner), zhodnocení výsledků a možnosti využití nově vydolovaných informací pro potřeby rozvoje podnikání.

Klíčová slova:

aukce, asociace, business inteligence, datový sklad, data mining, datová pumpa, elektronická komerce, elektronický obchod, ETL, Microsoft Business Intelligence Development Studio, OLAP, rapidMiner, rozhodovací strom, SQL, shlukování

Abstract and key words

Abstract:

This material describes problematic about e-commerce from different views. Theme orientation is business on Internet and its possibilities and extensions. Material contain overview of the most frequent e-commerce forms and the legal formalities related to this isme. The second part of the text describe more advanced possibilities how to make the business more effective. These are mainly problems of data warehousing and data mining in relation to Internet Business. The material contain description of data warehouse construction in Microsoft SQL Server 2005. It describes each sub-phase to the final reports and the use of newly acquired information. Further descripton of the most used data mining methods and their application in e-business environment with both commercial tools (Microsoft Business Intelligence Development Studio) and open source tools (RapidMiner). The final section of this part contain the result evaluation, usage possibilities of newly given information for business extension Leeds.

Key words:

Auction, association, business intelligence, clustering, data warehouse, data mining, data pump, decision tree, e-commerce, e-shop, ETL, Microsoft Business Intelligence Development Studio, OLAP, rapidMiner, SQL

Seznam použitých symbolů a zkratek

APEK	Asociace pro elektronickou komerci
a. s.	Akciová společnost
B2B	Business to business
B2C	Business to consumer
B2G	Business to government
C2B	Consumer to business
C2C	Consumer to consumer
BIDS	Business Intelligence Development Studio
BI	Business Intelligence
DM	Data mining
DS	Datový sklad
DSS	Decision support systém
ERP	Enterprice resource planning
ETL	Extract transform load
HOLAP	hybridní OLAP
k. s.	komanditní společnost
MS	Microsoft
MOLAP	multidimenzionální OLAP
OLTP	Online transaction processing
OLAP	Online analytical processing
ROLAP	relační OLAP
SOS	Sdružení na ochranu spotřebitelů
SAOP	Spotřebitelský audit obchodních podmínek
s. r. o.	Společnost s ručením omezeným
v. o. s.	Veřejná obchodní společnost
XML	Extensible markup language
XMLA	XML for analysis

Obsah

1 Úvod ... **1**

2 Elektronická komerce obecně ... **2**

2.1 Základní členění elektronické komerce podle subjektů ... 3

2.2 Důvody pro elektronické podnikání ... 4

2.2.1 Výhody a přínosy Internetu pro firmu ... 4

2.2.2 Nevýhody a rizika elektronické komerce ... 5

3 Legislativní požadavky ... **7**

3.1 Právní formy podnikání ... 7

3.2 Podnikání fyzických osob ... 8

3.3 Podnikání právnických osob ... 9

3.4 Nutné právní kroky na začátku podnikání ... 10

3.5 Jak vybrat správnou živnost pro daný podnikatelský záměr ... 12

3.6 Další legislativní omezení spojená s e-komercí ... 13

4 Formy elektronické komerce ... **15**

4.1 Elektronické obchody ... 15

4.1.1 Standardy a certifikace ... 15

4.2 Aukce, aukční portály ... 18

4.2.1 Další typy aukcí ... 19

4.3 Reklama na internetu ... 22

4.3.1 Základní druhy internetové reklamy ... 22

4.3.2 Možnosti plateb za reklamu ... 23

4.3.3 Problémy internetové reklamy ... 24

5 Úvod do problematiky Business Intelligence ... **25**

Obrázek č. 5: Hierarchie informačních úrovní ... 26

5.1 Přehled databázových systémů ... 26

5.2 Kvalita údajů pro analýzy ... 27

5.3 Nevhodnost transakčních databází pro analýzy ... 27

5.3.1 Decentralizovanost systémů OLTP ... 28

6 OLAP a datové sklady 30
6.1 Úvod do problematiky OLAP 30
6.1.1 Dimenze a fakta 30
6.1.2 Multidimenzionální databázový model 31
6.1.3 Úložiště multidimenzionálních dat 32
6.2 Datový sklad 33
6.3 Metody budování datového skladu 35
6.3.1 Metoda „velkého třesku“ 35
6.3.2 Přírůstková metoda 36
6.4 Etapa ETL 39
6.4.1 Extrakce 40
6.4.2 Transformace 40
6.4.3 Přenos 42

7 Vlastní řešení datového skladu 43
7.1 Zdrojová data 43
7.1.1 Popis zdrojových dat 44
7.1.2 Rozdělení atributů do dimenzionálních tabulek 46
7.1.3 Tabulky faktů 47
7.1.4 Hvězdicové schéma datového skladu 49
7.2 Architektura použitého řešení 50
7.2.1 Architektura analytických služeb MS SQL Serveru 2005 50
7.3 Integrace a zavádění dat (etapa ETL) 52
7.3.1 Integrační projekt DS 52
7.4 Vytvoření OLAP kostky 55
7.5 OLAP analýza a reporty 56
7.6 Zhodnocení výsledků OLAP analýzy 57
7.6.1 Shrnutí 66

8 Data Mining 67
8.1 Proces dobývání znalostí z databází 68
8.2 Předzpracování dat 69
8.2.1 Dělení dat z hlediska dolování znalostí 70
8.2.2 Filtrace dat 70

8.3 Algoritmy pro dolování znalostí z dat ... 71
8.3.1 Asociační pravidla ... 71
8.3.2 Rozhodovací stromy ... 74
8.3.3 Shlukování ... 75

9 Vlastní řešení data miningu ... 77
9.1 Předzpracování dat ... 77
9.2 Rozhodovací stromy ... 78
9.2.1 Dolování pomocí BIDS ... 78
9.2.2 Dolování pomocí RapidMineru ... 84
9.2.3 Zhodnocení výsledků ... 85
9.3 Asociace ... 86
9.3.1 Dolování prostřednictvím BIDS ... 87
9.3.2 Dolování pomocí RapidMineru ... 91
9.3.3 Zhodnocení výsledků ... 92
9.4 Shlukování ... 92
9.5 Porovnání BIDS a RapidMineru ... 97

10 Závěr ... 99

Seznam použité literatury ... 100
Internetové zdroje ... 100
Knižní zdroje ... 101

1 Úvod

Práci lze rozdělit na dva větší celky. První část práce se věnuje jednotlivým formám elektronické komerce a právním náležitostem, které jsou spojeny s provozováním obchodní činnosti na internetu. Na úvod je zařazena kapitola, která má za úkol uvést čtenáře do problematiky e-komerce, čili obecný popis problematiky, základní členění a jaké jsou hlavní důvody a z toho plynoucí výhody a rizika podnikání na internetu. Navazující kapitola se pak věnuje legislativním náležitostem souvisejícím s touto formou podnikání. Popisuje základní formy podnikání pro fyzické a právnické osoby, jak by měl subjekt postupovat v případě, že se rozhodne podnikat na internetu, v závěru tato část zohledňuje další právní náležitosti, které je potřeba mít na paměti. Následující kapitoly první části se pak již věnují popisu jednotlivých forem elektronického podnikání – aukční systémy, elektronické obchody a reklama na internetu. U každé z těchto forem jsou uvedeny pouze základní informace a teoretické aspekty. Detailnější popis, včetně příkladů lze poté nalézt v příloze textu. Příloha obsahuje výukový materiál s názvem *Podnikání na internetu.* V tomto výukovém materiálu je proveden detailnější rozbor této problematiky a měl by sloužit jako podpora při studiu pro studenty předmětu *Informační systémy pro elektronické podnikání.*

Druhá část textu se poté věnuje pokročilejším formám a možnostem, jak podnikání na internetu dále rozvíjet a využít tak již nabytých zkušeností a informací pro další rozvoj firmy. Možný je i druhý pohled a to, analýza trhu za využití těchto metod a následné strategické rozhodnutí o vstupu na trh, kdy toto rozhodnutí můžeme založit na informacích získaných z této analýzy a vytěžit tak určitou konkurenční výhodu. Jde tedy o problematiku datových skladů a data mingu aplikovanou do oblasti podnikání na internetu. Souhrnně se tato oblast nazývá *Business Itenlligence.*

Úvodní kapitola druhé části se věnuje obecnému úvodu do problematiky business intelligence, jaké jsou hlavní důvody k přechodu od transakčních databází k analytickým atd. Další kapitoly se poté věnují problematice datových skladů, je popsán celý proces budování datového skladu od prvotní analýzy zdrojových dat, přes jejich přípravu, datovou pumpu až po doručení výsledků konzumentům dat. Na závěr této části je shrnuto, jak nově nabyté informace využít pro další rozvoj podnikání společnosti. Datový sklad byl budován v prostředí Microsft SQL Server 2005 s využitím nástroje Microsoft Business Intelligence Developlment Studio (BIDS).

Závěrečná část práce se věnuje problematice dolování znalostí z dat – data miningu. Práce obsahuje popis základních metod používaných pro data mining a následně aplikaci těchto metod na zdrojová data. Pro potřeby data miningu byly použity jak komerční nástroje (BIDS), tak open source nástroje (RapidMiner) a to z toho důvodu, aby byly popsané postupy aplikovatelné napříč prostředími, které se mohou v jednotlivých společnostech vyskytovat. Na závěr popisu průběhu dolování prostřednictvím jednotlivých metod je popsáno, jak tyto získané informace využít v případě, že jsme v pozici, kdy přemýšlíme o podnikání v daném segmentu a provádíme analýzu trhu.

2 Elektronická komerce obecně

Pokud začneme na webu či v odborné literatuře vyhledávat pojem e-komerce, s největší pravděpodobností narazíme ještě na jeden pojem, a to *e-business*. Česky bychom mohli říct elektronické podnikání. Existuje však mezi oběma pojmy nějaký výrazný rozdíl? To není až tak jednoduché určit. Řada odborných publikací a článků používá tyto dva pojmy jako synonyma, řada jiných zase tyto pojmy odlišuje.
Např. v pojetí prof. Jaroslava Jandoše, CSc. jsou tyto dva pojmy vymezeny takto [14]:

Elektronickým obchodováním (e-komerce) rozumíme využívání informačních a komunikačních technologií v procesech prodeje a nákupu tj. v obchodní transakci.

E-podnikáním rozumíme mezipodnikovou integraci procesů, aplikací a systémů (založenou na využívání IS, tj ICT). Cílem je vyhovět měnícím se požadavkům zákazníků nabídkou nových mezipodnikových procesů, jakož i jejich integraci s novými podnikatelskými modely.

E-obchodování je tedy užší pojem než e-podnikání a zahrnuje pouze nákup a prodej zboží a služeb s využitím ICT. Obvykle ovšem platí, že pokud podniky využívají e-podnikání, pak také obvykle provozují e-obchodování.

Pro srovnání Bílá kniha o elektronickém obchodu uvádí [1]**:**
V Českém prostředí je samotný pojem „elektronický obchod“ vnímán buď ve smyslu veškerých obchodních aktivit, které zahrnují jak provozní, tak i technicko-logistické aktivity (e-business), nebo v užším smyslu, jeho obsahovou náplní je směna zboží a služeb za ekvivalentní hodnotu mezi jednotlivými prodávajícími a kupujícími, popř. zprostředkovateli (e-komerce).
V dalším textu jsou oba pojmy e-komerce a e-business chápány jako synonyma. Jejich odlišná definice je podle mého názoru důležitá zejména v oblasti ekonomiky. Z IT pohledu není nutné tyto pojmy nějak blíže rozlišovat.

Pro úplnost je uvedena ještě jedna komplexnější definice, která je široce přijímána také v mezinárodním prostředí. Jde o vymezení e-komerce podle organizace OECD - *Organisation for Economic Co-operation and Development* (Organizace pro hospodářskou spolupráci a rozvoj). Tato definice je rovněž základem mezinárodních statistik této organizace, Eurostatu i národních statistik členských zemí. Česká republika je členem této organizace od roku 1995.

Definice OECD vymezuje tři dimenze elektronického obchodování (má 3 části):

1. Podle použitých prostředků rozlišujeme tzv. širší a užší definici elektronického obchodování (elektronická transakce x internetová transakce)
2. Podle subjektů (zúčastněných stran) lze sestavit menší či větší matici druhů elektronického obchodování. Nejdůležitější je oblast *B2B*, následována *B2C*.
3. A konečně lze diskutovat, které (obchodní) procesy jsou do pojmu elektronický obchod zahrnuty.

2.1 Základní členění elektronické komerce podle subjektů

Nejčastěji se druhy elektronické komerce rozlišují podle subjektů (zúčastněných stran) na straně dodavatele a odběratele. Nejčastější je situace, kdy se rozlišují dva základní subjekty na každé straně (podnikatelé a spotřebitelé). V některých případech se přidává ještě subjekt třetí a to je vláda.
Při označovaní druhů elektronické komerce se používají vžité zkratky z angličtiny:

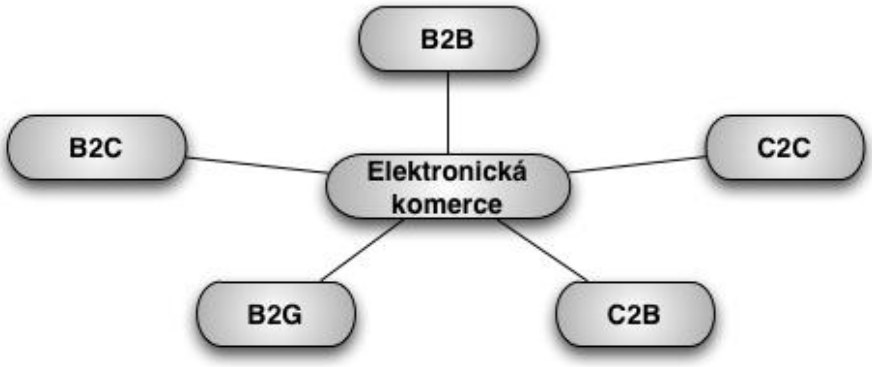
- **B2C (Business to Consumer)**
  jde o prodej zboží a služeb od podnikatelů (výrobců, obchodníků apod.) konečným spotřebitelům. V tomto případě se jedná o klasický obchod. V našem případě elektronický obchod.

- **B2B (Business to Business)**
  Prodej zboží a služeb mezi podnikatelskými subjekty, nejsou určeny ke konečné spotřebě. Díky využití moderních informačních technologií dochází k propojení všech transakcí v reálném čase, včetně finančních a logistických operací, urychlení a bezchybnosti ve firemním styku.

- **C2C (Consumer to Consumer)**
  Prodej zboží a služeb mezi spotřebiteli navzájem. Patří sem zejména různé aukce, spotřebitelská inzerce a další formy obchodu.

- **C2B (Consumer to Business)**
  Jde opět o prodej zboží a služeb koncovému spotřebiteli, ale rozdíl je v tom, že iniciativa zde vychází od spotřebitele, kdy konkrétní poptávka je umístěna někde na Internetu.

- **B2G (Business to Government)**
  Česky lze říci podnikání pro vládu. Jedná se o podmnožinu B2B a může být označována jako marketing veřejného sektoru. Jde o poskytování služeb a prodej produktů podnikatelských subjektů vládě (obecně statní správě). Organizace veřejného sektoru zadávají své požadavky (vyhlašují tendry) a potencionální zájemci na ně reagují (zasílají svou nabídku).



Obrázek č. 1: Základní členění e-komerce

2.2 Důvody pro elektronické podnikání

Důvody, proč se určitá společnost rozhodne podnikat na Internetu, budou u každé společnosti jiné – změna strategie firmy, expanze na nové trhy apod. Lze však zobecnit důvody proč vůbec o takovém kroku uvažovat. Jaké to společnosti může přinést konkurenční výhody, nové zákazníky atd. Tento krok má také svá rizika, proto i ty by měla být identifikována, konkretizována. Pokud bychom se nad tím hlouběji zamysleli, zjistíme, že jednotlivé důvody, výhody a rizika budou odlišná u zavedené společnosti, která bere Internet pouze jako další trh, kde se může realizovat. Na rozdíl od nově vznikající společnosti, která se rozhodně podnikat čistě pouze na Internetu.
Zaměříme se převážně na první skupinu. V závěru jsou poté popsány rozdíly, které mohou u obou těchto skupin nastat.

Internet jako nový trh
Pokud se společnost rozhodne expandovat na Internet, měla by již vědět, proč a jak chce Internet využívat, měla by si ujasnit cíle využití Internetu a jejich návaznost na celkovou strategii společnosti. Za další by měla být schopna si pro sebe odpovědět na následující otázky:

- Co je to Internet z pohledu dané firmy a co od něho mohu očekávat?
  - *Nový trh, prostředek pro komunikaci se zákazníkem, obchodními partnery atd.*
- Jaké nové možnosti mi Internet přináší?
  - *Urychlení podnikových procesů, nové skupiny zákazníků atd.*
- Jak Internet využít ve vlastním podnikání?
  - *Vybudujeme e-shop? Nebo využijeme pouze fenoménu sociálních stí pro podporu výrobků naší společnosti?*

Jedním z charakteristických rysů v počátcích elektronické komerce, ale u mnoha firem dodnes, je rozpor celkové strategie a cílů mimo Internet a na Internetu.
Pokud je společnost známá svým vřelým přístupem k zákazníkovi a najednou na Internetu nereaguje na dotazy návštěvníků svého elektronického obchodu, neinformuje je o aktuálních akcích, může tím poškodit své dobré jméno i mimo Internet a přijít tak o zákazníky, které již získala.

2.2.1 Výhody a přínosy Internetu pro firmu

Pokud firma dokáže plně integrovat Internet a jiné ICT zdroje, bude to ovlivňovat dosavadní stav v mnoha oblastech jejího fungování, mimo jiné:

- Celková strategie firmy
- Výrobky a služby
- Péče o zákazníky

Všechny tyto oblasti a mnohé další dostávají díky Internetu nový rozměr. Rázem existují nové možnosti jak např. pečovat o zákazníky.

Díky Internetu mohou mít třeba k dispozici ke každému zboží v obchodě recenze, diskuze, videa, na kterých si mohou prohlédnout dané zboží. Mohou on-line komunikovat s prodavačem, mají ihned k dispozici informaci o tom, zda jejich požadované zboží je k dispozici atd.

Hlavní přínosy elektronické komerce můžeme obecně rozdělit na:

- Finanční (tržby z prodeje)
- Nefinanční – např. zlepšení pověsti firmy.

Pokud bychom se zaměřili konkrétně na výhody, které Internet a elektronická komerce přináší, můžeme je rozdělit do několika skupin [14]:

Globálnost a permanentnost:

- Možnost oslovit globální trh za fixní náklady
- Informace a zboží jsou dostupné z celého světa
- Jsou dostupné 24 hodin denně
- Informace lze snadno aktualizovat
- Lze vybírat jen potřebné informace (možnost výběru má jak firma, tak zákazník)

Mnohostrannost využití:

- Lze využít text, obrázky, zvuky, video a animace
- Internet lze využít uvnitř firmy, při komunikaci s dodavateli, koncovými zákazníky

Efektivnost, úspora nákladů a času:

- Jednotná forma informací v elektronické podobě šetří náklady i čas
- Informace lze snadno aktualizovat, využít opakovaně, snáze archivovat a vyhledávat
- Rychlost a spolehlivost elektronické komunikace

Další výhody:

- Cennou zpětnovazební informaci lze získat mnoha způsoby
- Můžeme měřit účinnost nabídky a reakce zákazníků na ně
- Nové formy komunikace, reklamy a prodeje
- Nový zdroj informací
- Výrobky a služby lze distribuovat velmi rychle atd.

2.2.2 Nevýhody a rizika elektronické komerce

Samozřejmě elektronická komerce, kromě svých výhod, má také svá rizika a nevýhody, které je nutné znát a brát je v potaz. Po přehledu některých výhod, které Internet do světa podnikání přináší, následují nevýhody a rizika elektronické komerce. Opět je lze pro větší přehlednost rozdělit do několika skupin:

Problémy samotného Internetu

Toto je velice široká oblast, která zahrnuje jak technologické, tak etické či jiné problémy. Vyjmenujme alespoň některé:

- Přetíženost internetu – nárůst kapacit nestačí nárůstu uživatel a objemu dat
- Velký objem informací, kde se špatně hledá i s pomocí moderních vyhledávačů
- Snaha o zavedení sémantiky na web
- Bezpečnost uživatelů – viry, spyware, škodlivé kódy
- Spam

Rizika z pohledu zákazníka

- Obavy o soukromí a bezpečnost
  - Peněžní prostředky
  - Vlastní počítač
- Nedostatek vědomostí a zkušeností
  - Může vyvolávat i neoprávněné obavy
- Nemožnost fyzické prohlídky/vyzkoušení zboží
  - v ČR to řeší zákonná 14denní lhůta na možnost vrácení zboží bez udání důvodu v případě jeho nákupu v elektronickém obchodě – může docházet k zneužívání
- pro některé uživatele složitější úhrada nákupu

Rizika z pohledu firmy

- Nutnost osvojit si řadu nových technologií, postupů
- Riziko úniku informací, napadení WWW serveru, informačního systému firmy
- Svoboda Internetu sebou nese riziko zanesení nepravdivé, podvržené či zfalšované informace např. s cílem poškodit firmu

3 Legislativní požadavky

Elektronické podnikání se v základních principech neliší od podnikání klasického, proto je potřeba i v tomto případě stanovit právní rámec podnikání. Je potřeba učinit rozhodnutí zda podnikatel bude vystupovat jako fyzická či právnická osoba a následně vymezit konkrétní právní formu podnikání. Mimo to sebou elektronická komerce nese další právní náležitosti, které je nutné znát a respektovat. Této problematice se věnuje následující část textu. Rovněž obsahuje soupis právních činností, které je potřeba učinit při startu podnikání.

3.1 Právní formy podnikání

Právní forma podnikání specifikuje formu podnikání, pod kterou podnikatel vystupuje. Jednotlivé typy právních forem podnikání upravuje obchodní zákoník (obchodní společnosti a družstva). Živnostenský zákoník se pak zaměřuje na podnikatelské subjekty provozující svou činnost na základě živnostenského oprávnění. V České republice lze podnikat buď jako fyzická osoba nebo jako právnická osoba. Typy právních forem podnikání upravuje *obchodní zákoník v druhé části* [2] (obchodní společnosti a družstva – *Zákon č. 513/1991 Sb*.). Živnostenský zákon (*Zákon č. 455/1991 Sb*. [3]) se pak zaměřuje na podnikatelské subjekty provozující svou činnost na základě živnostenského oprávnění.

Každá právní forma podnikání má svá určitá specifika. Co může být výhodou u jedné, je nevýhodné u té druhé a obráceně. Proto si potenciální podnikatelé musí dobře rozmyslet, která z právních forem je pro jejich podnikání ta nejlepší s ohledem na různá kritéria. Také každý obor činnosti má svá specifika, každý člověk má jiné představy o svém podnikání, o budoucnosti a dalším směřování své firmy apod. Proto je vhodné si na začátku vše dobře rozmyslet, prostudovat si jednotlivé právní formy a případně se poradit s osobou znalou oboru. Na základě těchto poznatků poté učinit rozhodnutí jakou právní formu pro podnikání použít.

Jako **fyzická osoba** má možnost člověk podnikat na základě živnostenského oprávnění, které prokazuje výpisem ze živnostenského rejstříku, že může provozovat živnost:

- Ohlašovací – která se dále dělí na řemeslnou, vázanou a volnou
- Koncesovanou

Jinou možností je podnikat jako **právnická osoba**, což je uměle vytvořený subjekt, zapsaný do obchodního rejstříku, který podniká na území České republiky jedním z následujících způsobů:

- Společnost s ručením omezeným (s. r. o)
- Akciová společnost (a.s)
- Veřejná obchodní společnost (v.o.s)
- Komanditní společnost (k.s.)
- Družstvo.

V souvislosti s právními formami a s ohledem na následující část textu je vhodné na tomto místě uvést a vysvětlit základní pojmy z této oblasti [4]:

Podnik
Podnikem se rozumí soubor hmotných, jakožto i osobních a nehmotných složek podnikání. K podniku náleží věci, práva a jiné majetkové hodnoty, které patří podnikateli a slouží nebo mají sloužit k provozování podniku.

Obchodní majetek
Obchodním majetkem fyzické osoby je majetek (věci, pohledávky a jiná práva a penězi ocenitelné jiné hodnoty), který patří podnikateli a slouží nebo je určen k podnikání. Obchodním majetkem právnické osoby je její veškerý majetek.

Neoprávněné podnikání
Neoprávněné podnikání provozuje osoba, když uskutečňuje činnost (která vyžaduje ohlášení nebo povolení) bez ohlášení nebo povolení.

Místo podnikání
Místem podnikání fyzické osoby je adresa zapsaná jako místo podnikání v obchodním rejstříku nebo v jiné evidenci.

Obchodní firma
Obchodní firma je název, pod kterým je podnikatel zapsán v obchodním rejstříku. Tento pojem se tedy nevztahuje na fyzické osoby, které v obchodním rejstříku nejsou zapsány. Firmou fyzické osoby musí být vždy její jméno a příjmení.

3.2 Podnikání fyzických osob

V České republice je tato forma nejčastější formou podnikání. Využívají ji zejména ti, kdo ve svém podnikání nespatřují hlavní zdroj svých příjmů, ale není to pravidlem. Klidně se člověk podnikající prostřednictvím živnostenského oprávnění může touto formou živit. Zahájení u této formy je velice jednoduché. Podnikatel jako fyzická osoba vykonává nejčastěji činnost menšího rozsahu. Ve většině případů také podnik sám řídí a vede. Samozřejmě je možné ustanovit odpovědného zástupce.
Fyzickými osobami podle zákona jsou:

- Osoby podnikající na základě živnostenského oprávnění
- Osoby zapsané v obchodním rejstříku (zapisují se na vlastní žádost nebo povinně, poté co jim tato povinnost vznikla – tyto podmínky stanovuje obchodní zákoník v §34 – jestliže:
    - Výše jeho výnosů nebo příjmů (bez DPH) dosáhla nebo překročila za 2 po sobě bezprostředně následující účetní období v průměru částku 120 mil. Kč.
- Osoby podnikající na základě jiného oprávnění podle zvláštního předpisu (např. daňoví poradci, tlumočníci)
- Soukromě hospodařící zemědělci zapsaní v evidenci

Hlavní výhody této právní formy:

- Nízké výdaje na založení společnosti
- Podnikání lze ve většině případů zahájit ihned po ohlášení
- Samostatnost a volnost při rozhodování
- Snadné založení, přerušení nebo ukončení činnosti
- Místo podvojného účetnictví lze vést pouze daňovou evidenci
    - A to v případě pokud podnikatel není zapsán v obchodním rejstříku
    - Nebo mu nevyplynuly povinnosti přejít na podvojné účetnictví
- Není vkladová povinnost – zakládající osoba nemusí vložit na začátku žádnou částku, která by představovala základní kapitál společnosti
- Celý zisk po zdanění náleží podnikateli atd.

Kromě výhod má tato forma podnikání samozřejmě i své **nevýhody**. Zde jsou některé z nich:

- Vysoké riziko vyplývající z neomezeného ručení podnikatele za závazky společnosti
- Neomezený přístup k bankovním úvěrům, který může být spojený např. i s vysokou úrokovou mírou
- Vysoké nároky na odborné a ekonomické znalosti podnikatele – ten si v mnoha případech provádí veškerou administrativu sám
- Vzhledem ke své velikosti může v obchodních vztazích působit jako malý a nevýznamný partner.

3.3 Podnikání právnických osob

Podnikání právnických osob v České republice upravuje obchodní zákoník. Ten definuje a upravuje *obchodní společnosti* (založeny za účelem podnikání) a *družstva*.

Obchodní společnosti se rozdělují do 2 základních skupin podle toho, kolik osob danou společnost zakládá, jak tyto osoby ručí a zda mají povinnost vložit do společnosti vklad v určité výši [4]:

- **Osobní společnosti** – veřejná obchodní společnost (v. o. s.), komanditní společnost (k. s.)
    - Předpokládá se jejich osobní účast na řízení a neomezené ručení společníků za závazky společnosti

- **Kapitálové společnosti** – společnost s ručením omezeným (s. r. o.),
  akciová společnost (a. s.)
    - Jedná se o společnosti, ve kterých je povinností vložit vklad, který pak tvoří základní kapitál společnosti
    - Výše vkladu stanovuje zákon a je odlišná pro jednotlivé typy společností

- **Družstva** – společenství neuzavřeného počtu lidí (min. 5)
    - Základní kapitál je tvořen vklady
    - Společnost se zakládá za účelem podnikání nebo za účelem zajišťování potřeb svých členů

Všechny tyto subjekty musí být zapsány v obchodním rejstříku [5], což je veřejný seznam, ve kterém jsou evidovány subjekty, kterým to ukládá zákon.
Z celé této množiny právních forem pro právnické osoby bude blíže popsána pouze jediná a to společnost s ručením omezeným, jelikož je to hned po živnostnících nejčastější forma podnikání a nejrozšířenější podnikání právnických osob u nás.

Společnost s ručením omezeným (s. r. o.)

Společnost může být založena 1 osobou. Maximální počet společníků je 50. Základní kapitál společnosti tvoří vklady společníků, kteří ručí za závazky společnosti do výše souhrnu nesplacených částí vkladů všech společníků. Například pokud první společník má vklad splacen a druhý ne, ručí i ten první do výše nesplaceného vkladu druhého společníka. Jakmile všichni společníci splatí všechny vklady, již neručí. Výše základního kapitálu musí činit alespoň 200 000 Kč, přičemž každý společník musí vložit alespoň 20 000 Kč (viz §108 a §109 zákona č.513/1991 – obchodní zákoník). Vklady mohou být peněžité i nepeněžité (auto, nemovitost atd.). Společnost musí vytvářet rezervní fond. Mezi orgány společnosti patří valná hromada, coby nejvyšší orgán. Statutárním orgánem jsou jednatelé, kterým náleží obchodní vedení společnosti. Je možno stanovit i dozorčí radu.

Výhody:

- Společníci mají omezené ručení (jen do výše nesplacených vkladů)
- Není nutný souhlas všech společníků pro přijetí většiny rozhodnutí
- Vyplacené podíly na zisku společníkům (fyzickým osobám) nepodléhají pojistnému (sociální pojištění)
- Polovinu daně sražené z vyplácených podílů na zisku lze uplatnit jako slevu na dani společnosti

Nevýhody:

- Nutný počáteční kapitál (vklady)
- Dvojité zdanění – zisk společnosti je zdaněn daní z příjmů právnických osob a vyplacené podíly na zisku společníkům jsou dále zdaněny srážkovou daní – viz §36 zákona o daních z příjmů
- Administrativně náročnější založení a chod společnosti
    - Nutnost svolat valnou hromadu
    - Zápisy z valných hromad
    - Povinnost vést (podvojné) účetnictví

3.4 Nutné právní kroky na začátku podnikání

Následující text v bodech popisuje, jaké právní kroky je nezbytné učinit jednak při zahájení podnikání jako fyzická osoba a rovněž nutno dodržovat během vykonávání činnosti [6]:

- Nutno zjistit, zda je k podnikání potřeba nějaké oprávnění (živnostenské nebo jiné)
- Na živnostenském úřadě je potřeba provést registraci pro potřeby placení daní, pojištění. V případě nutnosti ohlásit živnost.
- Zjistit jaké pojištění a v jaké výši je nutné platit
- Povinností je archivovat všechny doklady, z nichž se na konci účetního období sestaví daňová evidence nebo evidence příjmů a výdajů
- Podání daňového přiznání a přehledy pojištění

Kdy je potřeba živnostenský list

Pokud osoba chce vydělávat samostatně, musí si zjistit, zda potřebuje:

- Živnostenský list
- Jiné oprávnění (daňoví poradci, tlumočníci)
- Nebo nic (autoři, spisovatelé, herci, hudebníci jsou takzvané Osoby samostatně výdělečně činné bez oprávnění)

Živnostenský úřad a registrace

Na Živnostenském úřadě žadatel vyplní jednotný registrační formulář. Tím provede najednou všechna potřebná ohlášení a registrace – pojišťovna, finanční i živnostenský úřad. Nyní již nepotřebuje předem zajišťovat výpis z rejstříku trestů, jak to bylo dříve. Dnes to za něj udělá úřad.

Pojištění

Podnikatel si musí zjistit, zda je povinen platit:

- Jen zdravotní pojištění nebo i sociální pojištění
- Měsíčně zálohy nebo stačí vše doplatit až po skončení roku po podání daňového přiznání

Pokud provozuje podnikání jako:

- Hlavní činnost – musí měsíčně platit zálohy na zdravotní i sociální pojištění
- Vedlejší činnost
    - Zdravotní a sociální pojištění doplatí na konci roku po podání daňového přiznaní. Pokud bude mít nízký zisk, sociální pojištění nebude muset platit vůbec.

Daňová a různé další evidence

Během roku se musí vést:

- Buď daňová evidence – tzn. je potřeba si schovávat všechny doklady
- Nebo evidence příjmů a výdajů
- Pokud se podnikatel zaregistroval jako plátce DPH, má rovněž povinnost vést evidenci podle zákona o DPH a to průběžně.
- Pokud uplatňuje výdaje na auto, musí vést evidenci jízd.

Daňová evidence nahrazuje dříve používaný pojem *jednoduché účetnictví*. Živnostník samozřejmě může dobrovolně vést také podvojné účetnictví, to však není nutné, pokud má podnikání jako vedlejší činnost nebo nedosahuje vysokého obratu a mnoha účetních případů během roku. Povinnost vést účetnictví ukládá zákon o účetnictví.

Daně

Po skončení roku musí podnikatel na finanční úřad podat daňové přiznání a zaplatit daň z příjmu. Pokud je plátce, musí navíc plnit veškeré povinnosti, které zákon ukládá plátcům DPH a po každém čtvrtletí podávat daňové přiznání k DPH. Pokud uplatňuje výdaje na auto, musí platit čtvrtletně zálohy na silniční daň a po konci roku podat daňové přiznání.

3.5 Jak vybrat správnou živnost pro daný podnikatelský záměr

Výše v textu je popsáno, že živnosti můžeme rozdělit na ohlašovací, které lze dále rozdělit na volné, vázané, řemeslné a koncesované. Nyní budou jednotlivé druhy živností popsány a poté bude ukázáno, jak postupovat při volbě vhodné živnosti pro určité podnikání.

Volné živnosti

Většina činností patří mezi volné živnosti. Čili pro získání živnostenského oprávnění žadateli stačí ohlásit se na příslušném úřadě a doložit splnění všeobecných podmínek. Mezi všeobecné podmínky patří minimální věk žadatele 18 let, způsobilost k právním úkonům a trestní bezúhonnost. Za živnostenské oprávnění žadatel zaplatí 1000 Kč.

Volná živnost tedy je multiprofesní živnost. Jde o souhrn činností pro výrobu, obchod a služby, které nevyžadují odbornou způsobilost. Mezi volné živnosti patří činnosti vyjmenované v příloze č. 4 živnostenského zákona (455/1991 Sb.).

Vázané živnosti

Pro získání vázané živnosti musí žadatel splňovat odbornou způsobilost, což může například být ukončené odborné vzdělání, rekvalifikace, nebo pokud je to požadováno i určitá délka praxe. Vše lze nalézt v příloze č. 2 živnostenského zákona. Opět zde platí stejné všeobecné podmínky jako v případě volné živnosti. Opět žadatel zaplatí 1000 Kč.

Řemeslné živnosti

Řemeslným živnostem se věnuje příloha č. 1 živnostenského zákona. Aby žadatel získal tuto živnost, musí opět splňovat odbornou způsobilost:

- Odborné vzdělání v příslušném oboru
- Odborné vzdělání v příbuzném oboru + rok praxe v oboru
- Byl určitý počet let OSVČ nebo zaměstnanec ve vedoucí funkci
- Má alespoň 6 let praxe v oboru

Rovněž i zde musí splnit všeobecné podmínky jako v minulých případech.

Koncesované živnosti

Koncesované živnosti lze nalézt v příloze č. 3 živnostenského zákona. Koncesovanou živnost žadatel neohlašuje, ale žádá o ni. Žádost podává na příslušném živnostenském úřadě podle místa bydliště. Opět musí splnit odbornou způsobilost např. požadované vzdělání. K dané žádosti se poté vyjadřuje příslušný orgán státní správy.

Jde tedy o regulovanou činnost, kdy uznávacím orgánem je Ministerstvo průmyslu a obchodu. Mezi koncesované živnosti např. patří taxislužby, pohřební služby či dražby.

Jakou živnost zvolit pro konkrétní podnikatelskou činnost

Při rozhodování, jaká živnost je pro dané podnikání ta nejlepší, je vhodné se inspirovat konkurencí. Čili zjistit, jakými oprávněními disponuje konkurence. Přitom je vhodné inspirovat se u zaběhnuté konkurence, která již funguje alespoň rok nebo i déle. To proto, aby alespoň jednou daná společnost prošla schvalovacím řízením a zaplatila daně. Takto lze lehce oddělit společnosti, které fungují dobře od těch méně úspěšných. Pro tyto účely dobře poslouží portál www.justice.cz, kde lze např. prohledávat obchodní rejstřík a zjistit si tak předmět podnikání konkurenčních společností. Další možností, jak dobře zvolit živnost pro podnikání, je nechat si poradit. Na každém živnostenském úřadě by měli být schopni podat adekvátní informace ohledně volby živnosti. Je vhodné se seznámit s uvedenými přílohami živnostenského zákona a snažit se najít činnost, která nejvíce vystihuje předmět daného podnikání.

3.6 Další legislativní omezení spojená s e-komercí

Kromě výše popsaných legislativních nařízení spojených s elektronickým podnikáním existuje ještě řada dalších právních nařízení, které je nutno respektovat. Tato nařízení již zpravidla souvisí s konkrétními technologickými řešeními pro elektronickou komerci (elektronické obchody, aukční systémy).

Autorské poplatky

10. 11. 2006 vstoupila v platnost vyhláška *488/2006 Sb.* k autorskému zákonu (*121/2000 Sb.*), která nově vyčísluje odměny autorům za prodej zboží, podléhající zpoplatnění dle autorského zákona. Proto je nutné u veškerého zboží, na něž se vztahují autorské odměny, uvést část ceny zboží odpovídající částce autorské odměny samostatně jako další položku na faktuře. Výše tohoto poplatku stanovuje zákon. Zboží, na které se tento poplatek vztahuje, je např. CD/DVD média, pevné disky a další paměťová média.

Recyklační poplatek

13. srpna 2005 vstoupila v platnost novela zákona o odpadech. Z novely vyplývají povinnosti pro výrobce a prodejce elektrospotřebičů, které se rovněž dotýkají konečných spotřebitelů:

Na základě zákona 185/2001 Sb. o elektroodpadech a jeho novely č. 7/2005 Sb. je prodejce zboží podléhajícího poplatkům (elektrozařízení) povinen kupujícímu od září 2005 vyúčtovat při prodeji náklady na likvidaci historických zařízení, tedy na ekologickou likvidaci. Výše těchto poplatků je stanovena sazebníkem sdružení dovozců a výrobců.

Čili prodejce je povinen informovat zákazníka o tom, že cena daného zboží obsahuje rovněž položku recyklační poplatek a tato částka by měla být na faktuře samostatně jako další položka. Rovněž by měl zákazníkovi umožnit odevzdat při koupi nového spotřebiče starší obdobného typu a nejen téže značky tak, aby mohla být zajištěna jeho ekologická likvidace. Tato povinnost je určena ustanovením §37, odst. 4 zákona o odpadech.

Další legislativní náležitosti související s elektronickými obchody

Provozování e-shopu je volnou živností. Podle přílohy č. 4 k zákonu *č. 455/1991 Sb*. (živnostenský zákon) spadá do oboru číslo 48 – Velkoobchod a maloobchod. Podnikání na internetu má rovněž svá pravidla. Podnikatel musí podnikat v souladu s platnými právními předpisy (například Směrnice Evropského parlamentu a Rady o elektronickém obchodu, Směrnice o prodeji na dálku, Občanský zákoník, zákon o ochraně spotřebitele). Elektronický obchod tak musí například povinně zveřejňovat informace o podnikateli nebo se při uzavírání smlouvy za použití prostředků komunikace na dálku řídit občanským zákoníkem. Podnikatel by si tak podle všech těchto norem měl formulovat své vlastní obchodní podmínky. Orientovat se v legislativně je však pro mnohé velice obtížné, a tak se dost často stává, že jednotliví podnikatelé opisují obchodní podmínky od sebe navzájem a to i s chybami. Pro zajištění správného jednání internetového obchodníka v souladu se všemi předpisy pak slouží certifikáty. Na českém trhu se uplatňují dva – SAOP a Certifikovaný obchod.

Spotřebitelský audit obchodních podmínek (SAOP) zaštiťuje Sdružení na ochranu spotřebitelů (SOS).

Certifikát SAOP o internetovém obchodníkovi říká, že jeho obchodní podmínky odpovídají zákonům a spotřebitel se nemusí obávat svého zneužití. Certifikát platí pouze jeden rok. Poté je nutné jeho obnovení.

Certifikovaný obchod je certifikát organizace APEK (Asociace pro elektronickou komerci) a detailněji je popsán v kapitole o elektronických obchodech.

Legislativa spojená s aukcemi

Pojem elektronická aukce je definován v zákoně o veřejných zakázkách (§ 96 a § 97 zákona číslo *137/2006 Sb.)*. Elektronickou aukci je možné realizovat pouze za použití elektronických prostředků. Detailní informace o elektronických aukcích a jejich náležitostech lze nalézt v uvedeném zákoně.

Provozování internetové aukce není upraveno žádným speciálním zákonem, jde tedy o běžnou podnikatelskou aktivitu. Provozovatel aukce je stejně jako provozovatel elektronického obchodu povinen vydat Všeobecné obchodní podmínky v souladu s ustanovením § 273 zákona *č.531/1991 Sb.*, obchodního zákoníku. Všeobecné obchodní podmínky jsou nedílnou součásti Smlouvy o poskytování služeb, která se uzavírá mezi provozovatelem aukčního systému (elektronického obchodu) a právnickými či fyzickými osobami, které využívají služeb provozovatele, který tyto služby poskytuje právě na základně dané smlouvy. Všeobecné obchodní podmínky se vztahují na všechny smluvní vztahy uzavírané mezi provozovatelem a jeho zákazníky.

4 Formy elektronické komerce

Tato kapitola popisuje některé z běžných a pravděpodobně také nejpoužívanějších forem elektronické komerce, pomocí kterých lze na internetu podnikat. Jde o elektronické obchody a aukční portály. Tyto dvě formy pak doplňuje část textu věnovaná reklamě na internetu, ta sama o sobě nepředstavuje přímo formu elektronické komerce, nicméně s podnikáním velice úzce souvisí a bez kvalitní a dobře cílené reklamy se dnes na Internetu podniká velice špatně.

4.1 Elektronické obchody

První internetové obchody se objevily v USA již v první polovině 90. let 20. století. Bouřlivý rozvoj však zaznamenaly až po roce 2000. V současné době nabízejí široké spektrum zboží i služeb s využitím pokročilých způsobů plateb a stávají se alternativou kamenného obchodu nebo nákupního centra.
Rozvoj elektronického obchodu v České republice se datuje od roku 1996, kdy byly založeny první internetové obchody. Celkový obrat internetového obchodování v České republice (spotřebitelské nákupy) byl odhadnut na 22 miliard korun v roce 2008, což je cca 25% nárůst proti předchozímu roku (v odhadu není zahrnuto cestování a zábava). V roce 2008 mělo s nákupem na Internetu podle odhadů zkušenost okolo 2 milionů obyvatel ČR.
Rozvoj moderních technologií a neustálé rozšiřování internetu má za následek, že elektronické obchody jsou v dnešní době již pro mnoho významných společností hlavním předmětem jejich podnikání a pro další nemalou skupinu představují e-shopy nezanedbatelnou položku v jejich příjmech.

4.1.1 Standardy a certifikace

Jelikož již v dnešní době, existuje velké množství elektronických obchodů a ne všechny se dají považovat za seriozní, existují organizace, které se mimo jiné zabývají certifikací elektronických obchodů. Jde jednak o službu spotřebitelům, kdy takto certifikovaný obchod lze považovat za důvěryhodný, poskytující větší záruky a jistotu při nákupu. Na druhou stranu certifikace může být dobrým signálem také pro obchodní partnery. V české republice se certifikaci v oblasti internetových obchodů věnuje organizace **APEK – Asociace pro elektronickou komerci:**

- Jde o sdružení více jak 150 firem, podnikatelů a odborníků v elektronickém obchodu. Asociace byla založena v roce 1998 jako nezávislá organizace, která podporuje rozvoj elektronického obchodu v České republice. Mezi členy APEKu patří největší české internetové obchody, přední softwarové společnosti a finanční instituce.

Asociace se zaměřuje především na služby pro své členy:

- Analýzy a studie o elektronickém obchodu
- Vytváření a podpora etických principů podnikání
- Workshopy, semináře, vzdělávání

Certifikace APEK

a) APEK Certifikovaný obchod [7]

Jde o novou certifikaci organizace, která nahrazuje starší – *Nákup bez obav*. *Certifikace APEK Certifikovaný obchod* zaručuje zákazníkům internetových obchodů, že certifikovaný obchodník splňuje základní pravidla bezpečného a bezproblémového nákupu, jejichž úroveň je stanovena certifikačními pravidly. Obchodník mimo jiné dodržuje:

- Úplné a pravdivé informování o provozovateli (sídlo obchodníka, kontakty na odpovědné osoby, apod.)
- Úplné a pravdivé informování o zboží a cenách, včetně všech poplatků

Při certifikaci se dále rovněž kontroluje:

- Jakým způsobem probíhá nákup (nákupní řád)
- Jak probíhá reklamace (reklamační řád)
- Komunikace se zákazníky (odpovídá na e-maily, telefonáty, ap.)
- Splňuje zákonné požadavky, dané zejména směrnicemi Evropského parlamentu a Rady, občanským zákoníkem a dalšími normami

Obrázek č. 2: Logo organizace APEK pro certifikovaný obchod

Co je to certifikace?

Certifikace je proces hodnocení internetového obchodu podle Certifikačních pravidel, jehož úspěšným završením je vydání certifikátu. Stvrzuje, že certifikovaný obchod dodržuje základní pravidla bezpečného a bezproblémového nákupu, zejména úplné a pravdivé informování o provozovateli, procesu nákupu, vyřízení objednávky a reklamacích a bezproblémovou komunikaci se zákazníkem. Certifikační pravidla stanovuje Asociace pro elektronickou komerci jako tuzemská autorita v oblasti elektronického obchodu. Certifikace provádí APEK již od roku 1999.

b) Certifikace II. Stupně

Proces certifikace druhého stupně může podstoupit každý nositel značky *APEK - Certifikovaný obchod*. E-shop je v několika kolech testován pomocí metody "*Mystery-shopping*"[9].

Hlavním cílem této certifikace je pomoci zákazníkovi rozpoznat obchody disponující velkou kvalitou služeb. Zároveň tento test pomáhá obchodníkovi identifikovat oblasti pro možná zlepšení nabízených služeb.

V rámci Certifikace II. stupně je testováno 6 oblastí (Komunikace a kontakty, Informace o výrobcích, Kvalita nákupního řádu, Reklamace a odstoupení od smlouvy, Existence a kvalita nadstandardních služeb, Objednání a doručení zboží). Velký důraz je kladen především na oblasti testované přímou metodou nákupu.

Testované oblasti v rámci druhého stupně certifikace APEK:

Komunikace a kontakty

- Telefonická komunikace
- Dostupnost na telefonu
- Zvýhodněné tarify
- Kvalita telefonické komunikace
- E-mailová komunikace
- Rychlost odezvy
- Kvalita emailové komunikace

Informace o výrobcích

- Zvýraznění konečné ceny
- Prezentace výrobku
- Kodex terminologie lhůt dodání

Kvalita nákupního řádu

- Držitel SAOP (Spotřebitelský audit obchodních podmínek)

Reklamace a odstoupení od smlouvy

- Odstoupení v zákonné lhůtě 14 dní
- Vrácení peněz
- Reakce obchodníka na žádost o vrácení
- Reklamace
- Přijetí reklamace
- Vyřízení reklamace

Existence a kvalita nadstandardních služeb

- Kamenná prodejna
    - Existence a lokace kamenných poboček
    - Otvírací doba
    - Nabízené služby
- Možnost placení
- Bezbariérová přístupnost webu

Objednání a doručení zboží

- Objednávka
    - Komunikace obchodníka

- Doručení skladového zboží
    - Proces doručování
    - Dokumenty

- Doručení neskladového zboží
    - Expedice
    - Proces doručování
    - Dokumenty

Obrázek č. 3: Logo organizace APEK pro certifikovaný obchod 2. stupně

c) Kodex terminologie lhůt dodání [8]

APEK také představil „*Kodex terminologie lhůt dodání zboží u internetových obchodníků*." Obchodníci hlásící se k tomuto kodexu používají na svých stránkách pouze jednoznačné a nezavádějící údaje.

Příklad:
U zboží je jako lhůta dodání uvedeno *ihned k odběru*: V případě existence kamenné prodejny/výdejny musí být zboží okamžitě dostupné k odběru na prodejně! Pakliže tomu tak není, je nutné u lhůty dodání použít odpovídající časový údaj.

4.2 Aukce, aukční portály

Aukce, aukční portály jsou dalším řešením po elektronických obchodech, které se používají pro e-komerci. V zásadě jsou si obě řešení velice podobná po technické stránce. Hlavní rozdíl je v modelu, který se používá pro prodej zboží.

Aukce lze definovat například takto:
Aukce (též dražba) je zvláštní forma trhu, kde se soustřeďuje poptávka a nabídka po konkrétním druhu zboží. Většinou se aukce zúčastňuje malý počet prodávajících a větší počet kupujících.
Jedná se o veřejnou dražbu zboží, které nemá ani u stejného druhu nebo sortimentu stejnou kvalitu, nebo stejné technické parametry, nebo jakost.

Předmětem aukcí může být prakticky cokoliv, co má nějakou hodnotu a prodávající si myslí, že o dané zboží bude zájem. Přece ale existují některé typické komodity:

- starožitnosti a umělecké předměty,
- nemovitosti a parcely,
- elektronika.

Cena zboží se vyhlašuje jako cena nejnižší (tzv. anglická aukce) nebo nejvyšší (tzv. holandská aukce). Obchod je uzavřen přiklepnutím zboží kupujícímu. Aukce jsou buď *pravidelné* (např. exekutorský úřad pravidelně draží předměty zabavené při exekucích), nebo *příležitostné* (např. aukce uživatelů na aukčních portálech).

Průběh aukcí v prostředí Internetu

V prostředí internetu, na běžných aukčních portálech jako *aukro.cz, ikup.cz, odklepnuto.cz* či *ebay.com*. Aukce probíhají takovým způsobem, že prodávající vystaví na portálu své zboží, definuje vyvolávací cenu, dobu trvání aukce. Kupující, kteří mají o dané zboží zájem, nabízí takovou částku, která je vyšší než aktuální prodejní cena. Zboží získává ten kupující, který nabídne nejvíc. Toto je základní model aukcí, existuje celá řada jeho modifikací, jako např. po každém příhozu se prodlouží doba do konce aukce a nějaký předem stanovený čas aj. Jiný model aukcí je velice populární v obchodním prostředí. Společnost potřebuje nového dodavatele pro nějakou službu či výrobek potřebný ke svému podnikání. Může se jednat o software na zakázku, stejně tak třeba o novou výrobní linku. V tomto případě firma vyhlásí aukci a to tak, že definuje maximální možnou cenu, kterou je ochotna za danou službu či výrobek zaplatit. Aukci a tedy i kontrakt se zadávající společností. Vyhrává ten, kdo nabídne nejnižší cenu. Mimo tyto celkem běžné úpravy aukčního modelu prodeje zboží, existují také modifikace, které lze označit za neetické (např. aukce s platbou za příhoz), tento typ aukce bude popsán později v této kapitole.

Aukční portál

Aukční portál je webová aplikace, která nabízí svým návštěvníkům prodej a nákup zboží. Mezi nejznámější aukční portály můžeme zařadit *aukro.cz* a *ebay.com*.
Tyto velice úspěšné aukční portály dnes nabízí celou řadu doplňkových služeb pro své návštěvníky, jako je např. ochrana proti nedůvěryhodným prodejcům, v mnoha případech řeší reklamace zboží, nabízí vlastní platební systém, který se již stal standardem v oblasti plateb přes internet (payPall) atd. Kromě těchto velice známých portálů a dnes již také značně úspěšných firem existují další portály, které mohou být specializovány pouze na určitou komoditu či uzavřeny pouze pro určitou skupinu uživatelů. Dnes již je běžnou praxí také to, že velké společnosti nebo vládní organizace využívají různých specializovaných aukčních systémů pro nákup nových řešení pro svůj business.

Obrázek č. 4: Logo aukčního portálu ebay.com

4.2.1 Další typy aukcí

Kromě dříve popsaného obchodního modelu aukčních systémů (je stanovena vyvolávací cena a uživatele zasílají své nabídky), existují druhy aukcí, která staví na jiném obchodním modelu a některé z nich se dokonce pohybují na hraně zákona.

Reverzní aukce

Tento typ aukce se nejčastěji používá ve státní správě (přidělování veřejných zakázek) nebo v podnikové sféře (různá výběrová řízení). Jelikož každý podnik pro svou činnost nakupuje zboží a služby od různých dodavatelů. Finanční prostředky, které podnik na tyto položky vynakládá, tvoří podstatnou část veškerých jejich nákladů. Proto je snaha firem nakupovat za co nejnižší ceny. Nejnižší cenu může zajistit online výběrové řízení s použitím reverzní aukce.

Způsob fungování reverzních aukcí

Zadavatel stanoví maximální cenu, kterou je ochoten za zboží, službu či veřejnou zakázku investovat. Potencionální dodavatelé, kteří mají o danou zakázku zájem, nabízí nižší cenu než je udaná maximální cena zadavatele. Vítězí ten, kdo nabídne nejnižší cenu.

Fungování reverzních aukcí v prostředí Internetu

Zadavatel vloží zakázku do aukčního systému se všemi jejími náležitostmi a informacemi, které by měli potencionální dodavatelé mít k dispozici (maximální cena, termín zhotovení, požadavky na provedení atd.). Rovněž specifikuje požadavky na dodavatele. Zda je nutné, aby daný dodavatel byl držitelem některého z ISO certifikátů apod. Dodavatelé by se pak měli prezentovat také svými referencemi, které mohou zvýšit naději na úspěch atd. Je stanoven čas, kdy aukce začíná. Aukce bývají do systému vloženy s dostatečným předstihem před skutečným začátkem licitování, aby si potencionální dodavatelé mohli dostatečně rozmyslet a zhodnotit, zda mají o tuto zakázku skutečně zájem. V momentě, kdy aukce skutečně začíná, se dodavatelé přihlásí do systému a vkládají své nabídky. Samozřejmě mají k dispozici informace, na kterém místě se zrovna jejich nabídka nachází, jaká je minimální nabídka atd. Samotná aukce většinou trvá 30minut, v případě delšího boje je čas ukončení aukce automaticky prodlužován.

Přínosy tohoto systému

Hlavním přínosem je jednoznačně úspora nákladů a to na obou stranách – jak zadavatele, tak dodavatele. V případě zadavatele je hlavní přínos v tom, že získá dodavatele pro svou zakázku za prakticky nejnižší cenu, za kterou jsou ochotny firmy tuto zakázku realizovat.
Další aspekt je úspora nákladů na přípravu. Zde dochází k úspoře na obou stranách. Účastnici aukce nemusí být v daný okamžik v jedné místnosti, ale stačí pouze počítač připojený k Internetu.

Aukce s fiktivními uživateli, platbou za příhoz, sms aukce

Všechny položky, které jsou zmíněny v nadpisu tohoto odstavce, mají jedno společné. Nejde o 3 druhy aukcí, ale v naprosté většině případů se všechny tyto možnosti vyskytují najednou. Společným rysem aukčního portálu, který staví právě na těchto principech je to, že prakticky balancují na hraně zákona. Mnoho lidí si myslí, že v tomto případě se již jedná o hazard či loterii a nikoliv o aukční portál. Někdy jsou označovány jako tzv. *penny aukce*.

Základní princip takovéto aukce

Provozovatel webu vystaví nějaký lukrativní předmět (televize, mobilní telefon, notebook atd.) s nulovou vyvolávací cenou a odpočtem zpravidla na 24 hodin. Jeden příhoz zvýší cenu zboží o jednu jednotku – zpravidla to bývá jeden halíř nebo podobně nízká částka. S příhozem se rovněž prodlouží doba trvání aukce o 30 sekund. V čem je tedy ten háček?

Nepřihazují se skutečné peníze, ale od provozovatele portálu se nakupují příhozy. Nejčastěji stojí příhoz 5 Kč, takže např. člověk nakoupí za 200 Kč 40 kreditů (příhozů).
Za každý příhoz se tedy zvedne hodnota prodávaného předmětu o haléř a z konta přihazujícího se odečte 5 Kč (1 kredit). Ta pravá dražba začíná až v poslední minutě, jelikož úkolem všech je přihodit jako poslední před uplynutím intervalu, který se ovšem pořád obnovuje díky příhozům dalších dražitelů. Čili je velice obtížné trefit se do správného intervalu.

Nyní malý příklad a výpočet:

V nabídce bude fotoaparát v ceně 20 000 Kč aktuální cenou 230 Kč což je 23 000 halířů x 5 Kč = 115 000 Kč – 20 000 Kč (skutečná cena fotoaparátu) = 95 000 Kč zisk pro majitele aukce, která ještě nekončí. Prodejce si tedy klidně může dovolit prodat tento fotoaparát třeba za 250 Kč.
Ano, je zde nějaká malá šance na získání zajímavého předmětu za zajímavou cenu, pokud zrovna má dotyčná osoba štěstí a nevložila do dané aukce např. 10 000 Kč na příhozech (udělal to někdo jiný). Dalo by se to přirovnat k ruletě nebo výhernímu automatu.

Přihazovat lze i prostřednictvím sms. Odtud tedy pojem *sms aukce*, v tomto případě bývá však cena příhozu vyšší než 5 Kč. Do celého procesu však vstupuje ještě jeden faktor a to jsou fiktivní uživatelé (roboti). Pokud je aukce nová nebo má málo uživatelů nastupují roboti, kteří plní funkci fiktivních uživatelů a uměle navyšují cenu. V mnoha případech, kdyby provozovatel takovéhoto aukčního portálu neměl k dispozici (*virtuální uživatelé*) roboty, zcela jistě by dříve nebo později dospěl ke krachu, protože aktivní uživatelé by pouze párkrát přihodili a získali by například zmiňovaný fotoaparát za 10 Kč. Roboti ovšem nepřinášejí očekávaný příjem za příhoz, pouze uměle navyšují cenu.

Jak identifikovat roboty

Roboty lze identifikovat vcelku snadno. Stačí pouze chvíli sledovat některé aukce. Většinou se účastní aukcí, které „již hoří“, tzn. je půl minuta do konce a cena za zboží je směšně nízká. V takovém případě se zde odehrává souboj dvou uživatelů (robotů). Soupeřit s roboty je ve většině případů nerovný souboj, jelikož může být nastaven tak, aby přihodil v poslední setině sekundy před koncem aukce. Toto lze chápat jako podvodné jednání a nekalou soutěž, což je postižitelné zákonem.

Tyto aukce se pohybují na hraně zákona, jelikož souvisí s hazardem. Ve svém principu se příliš neliší od výherních automatů – vhodí se mince (příhoz v aukci) a možná daná osoba vyhraje. To se stane až v momentě, kdy je automat dostatečně „nakrmen” = prodejci se vrátily počáteční náklady a má dostatečný zisk. Čili šance zde získat zboží za výhodné ceny existuje, je však založena na náhodě a štěstí.
V českém prostředí na tomto principu fungují např. portály *bonus.cz*, dále pak *cosmito.cz* nebo *silena-aukce.cz*

4.3 Reklama na internetu

Za počátek systematického využití Internetu jako reklamního média se většinou považuje rok 1994. První reklamní proužky (bannery) se objevily na portále *Yahoo!* a webu počítačového magazínu *Wired*. Prozíraví lidé velice brzy odhalili výhody Internetu jako reklamního média.

Mezi tyto výhody můžeme zařadit rychlou odezvu, relativně přesné zacílení na určitou skupinu uživatelů a snadná analýza výsledných reklamních kampaní.

V České republice se o rozvoj internetové reklamy zasloužil především portál *Seznam.cz*, na kterém se první bannery začaly objevovat v roce 1996. Dále například portál *Billboard.cz*.

Internetová reklama od svého vzniku pochází celkem bouřlivým vývojem. Neustále roste objem reklamy, rovněž se mění její formy a přibývají nové.

Mediální agentura OMD [10] odhadla objem internetové reklamy v ČR na 710 milionů v roce 2004. Pro rok 2005 se uváděl odhad, kdy měl objem internetové reklamy poprvé v historii překročit 1 miliardu korun. V roce 2008 již tuzemští inzerenti utratili na internetové reklamě 5 miliard korun. V roce 2005 patřila internetové reklamě až pátá příčka co se týče objemu peněz investovaných do reklamy obecně. Její tržní podíl činil 5% a před ní se umístila televize (43,7 %), tisková reklama (33 %), rozhlas (10,2%), venkovní reklama (7,7 %). V roce 2008 však internetové reklamě patří třetí příčka hned za televizí a rozhlasem. Její tržní podíl již činí 9%. Z těchto pár čísel je patrné, jak dynamicky se rozvíjí internetová reklama a jak rychle roste. Oproti jiným typům reklamy (TV, rozhlas, tisk) zaznamenává nejrychlejší růst. S tím souvisí i neustále se zvyšující cena internetové reklamy. Odhady do budoucna hovoří o neustále rostoucím podílu internetové reklamy na úkor jiných médií. Proto je vhodné věnovat i tomuto odvětví e-komerce patřičnou pozornost a snažit se z ní vytěžit maximum.

4.3.1 Základní druhy internetové reklamy

Díky vývoji Internetu, reklamního trhu a e-komerce vznikly v průběhu let desítky forem internetové reklamy. Rovněž existuje mnoho způsobů, jak internetovou reklamu členit. Členění, které je uvedeno zde, je poměrně jednoduché a přehledné (má pouze čtyři kategorie). Je však velice pravděpodobné, že v budoucnu může dojít k doplnění dalších kategorií [14].

Základní druhy internetové reklamy:

1. E-mailová reklama
2. Grafická reklama
    a) Reklamní proužky neboli bannery
    b) Pop-up okna
3. Textová reklama
    a) „obyčejná“ textová reklama (nekontextová)
    b) Kontextová textová reklama (vázaná na klíčová slova, kontext článků)
    c) Firemní systémy textové reklamy (Seznam, Google)

4. Ostatní formy
    a) Virální marketing
    b) Partnerské programy
    c) Nepřímá reklama
    d) Sponzorování obsahu
    a další.

4.3.2 Možnosti plateb za reklamu

Tato část popisuje základní a v současnosti rovněž nejpoužívanější modely způsobu plateb za reklamu.

a) Časová cena (paušální model)
Tento způsob platby se používal hlavně v počátcích grafické (proužkové reklamy). V současné době ovšem u tohoto druhu reklamy převažuje model platby podle návštěvnosti. Paušální cena je nyní typická např. u některých forem textové reklamy na portálech jako jsou *Seznam.cz, Centrum.cz* apod.

Výhody:
- Jednoduchost určení ceny za reklamu jak z hlediska provozovatele serveru, tak i inzerenta
- Reklama je pevně umístěna
- Předem známá koncová cena za reklamní kampaň

b) Dle návštěvnosti (podle počtu zobrazení)
Jde o typický model pro grafickou reklamu v současnosti. Udává se cena za 1000 zobrazení. Tento model je běžně označován zkratkou *CPM: cost per mile*. Díky dostupnosti podrobných statistik návštěvnosti jednotlivých serverů za dlouhodobé období se tento model v průběhu doby stal výhodnější, než model časový a to pro obě strany (inzerent, provozovatel serveru). Dnes jsou již dobře známy také **nevýhody** tohoto modelu:
- Nemotivuje reklamní agenturu ke kvalitě
- Vzrůstající tzv. „bannerová slepota“ u uživatelů
- Počet zobrazení neříká nic o tom, jak reklama na uživatele působí (např. zda na jejím základě koupí příslušný výrobek)
- Neúspěšné snahy motivovat zákazníky a platit jim za sledování reklamy

c) Pay-per-click (PPC)
Tento způsob je spjat hlavně s rozvojem kontextové reklamy. Typickým představitelem *je Google AdWord/AdSense*. Ani tento způsob nezaručuje inzerentovi, že si zákazník na základě reklamy dané zboží koupí. Avšak díky tomu, jak je celý systém konstruován, lze přinejmenším očekávat, že pokud zákazník na reklamu klikne, aspoň trochu jej zajímá. Tento model zažívá v poslední době nebývalý růst.

4.3.3 Problémy internetové reklamy

Internetová reklama se v dnešní době potýká také s několika problémy, které trápí jak inzerenty, tak uživatele internetu. Jedním z těchto problémů je tzv. fenomén reklamní slepoty.

Reklamní slepota (někdy také označována *bannerová slepota*) trápí zejména poskytovatele a inzerenty, kteří využívají grafickou, v poslední době však i kontextovou reklamu. Uživatelé internetu mají zažitý způsob, jakým na stránku (její obsah) nahlíží. Naučili se vnímat pouze obsah stránky a veškeré elementy, které se vyskytují okolo, prostě ignorují. Proto je v dnešní době míra prokliků u grafické reklamy v řádech promilí. Určitým způsobem, jak tento problém řešit představovala kontextová reklama, která se snažila nabídnout reklamu, svázanou s obsahem dané stránky. Ovšem boxy s kontextovou reklamou se také vyskytují různě po stranách, takže v mnoha případech to může dopadnout stejně jako v minulém případě. Nicméně úspěšnost této reklamy je podstatně vyšší než v případě grafické reklamy (bannery). Z tohoto důvodu se poskytovatelé internetové reklamy snaží reklamu dělat ještě agresivnější a donutit tak uživatele na ni kliknout. Proto přes stránky létají otravné objekty nebo na nás dokonce vyskočí nějaké vyskakovací okno, přehraje se reklamní video, které překrývá obsah atd. A poskytovatelé reklamního prostoru to inzerentovi mnohdy rádi dovolí, protože na penězích z reklamy jsou životně závislí. To však v lidech pěstuje ještě větší odpor vůči takto agresivní internetové reklamě. Takže problém se tím jen zhoršuje.

Druhým značným problémem internetové reklamy je SPAM, neboli nevyžádaná elektronická pošta. Tento fenomén plní e-mailové schránky každého uživatele elektronické pošty. SPAM je dílem robotů, kteří různě po internetu, mnohdy protiprávně sbírají e-mailové adresy a následně na ně zasílají velké množství reklamních e-mailů. V dnešní době však naprostá většina poskytovatelů e-mailových služeb nabízí tzv. spam filtr, který tyto nežádoucí zprávy zachytává a k uživateli se tak ve většině případů vůbec nedostanou. Podobné praktiky také používají veškeré společnosti, které tak chrání své zaměstnance a firemní komunikaci. Rozesílání SPAMu je dnes již trestné.

V České republice od 7. září 2004 platí *Zákon o některých službách informační společnosti* (č. 480/2004 Sb.), který problematiku SPAMu upravuje a vyžaduje prokazatelný souhlas příjemce zprávy. Dohledem nad dodržováním zákona byl pověřen *Úřad pro ochranu osobních údajů*. Tento zákon byl postupně novelizován. Zákon byl vytvořen podle směrnice Evropského společenství č. 2000/31/ES. SPAM definuje jako obchodní sdělení, což jsou všechny formy sdělení určeného k přímé či nepřímé podpoře zboží či služeb.

Obecně problém SPAMu jednoznačně snížil důvěryhodnost e-mailové reklamy a představuje jeden z největších problémů dnešního Internetu.

5 Úvod do problematiky Business Intelligence

V současnosti existuje již velice málo firem, které by nedisponovaly nějakým informačním systémem. Je však podstatně méně firem, které *data* získaná pomocí těchto informačních systémů efektivně využívají. Stále více společností, řeší otázku pořízení odpovídajícího řešení pro manažerské rozhodování (Business Intelligence – BI), které jim umožní práci s *informacemi*, které jsou v daný okamžik podstatné pro okamžité a správné rozhodnutí. Systémy BI jsou nápomocny při řešení problémů informačních a transakčních systémů, zároveň pomáhají zkvalitnit a zefektivnit řízen společnosti. Systémy BI představují určitou strategii práce s informacemi, jsou založeny na dodání informací do správných rukou, ve správný čas, v požadovaném formátu a v nezkreslené podobě.

Business Intelligence představuje ve zkratce pomoc uživatelům na všech úrovních organizace získat potřebné informace, aby se dokázali rychleji a lépe rozhodovat. Tato skutečnost může významně přispívat k rozvoji celého podniku. V dnešní době již BI není doménou pouze vedoucích pracovníků a analytiků. Každý pracovník společnosti může získat přístup k důležitým podnikovým údajům, které mu mohou pomoci při strategických, ale i běžných rozhodnutích. Díky tomu, že mají pracovníci přístup k informacím ve vhodnou dobu, mohou provádět důležité činnosti v souladu se strategií společnosti.

Business Intelligence pracuje s dvěma zdánlivě zaměnitelnými pojmy – *data* a *informace*. Ovšem tyto pojmy jsou zaměnitelné pouze na první pohled. Data se stávají informacemi, pokud:

- máme data,
- víme, že máme data,
- víme, kde tato data máme,
- máme k nim přístup,
- zdroji dat můžeme důvěřovat.

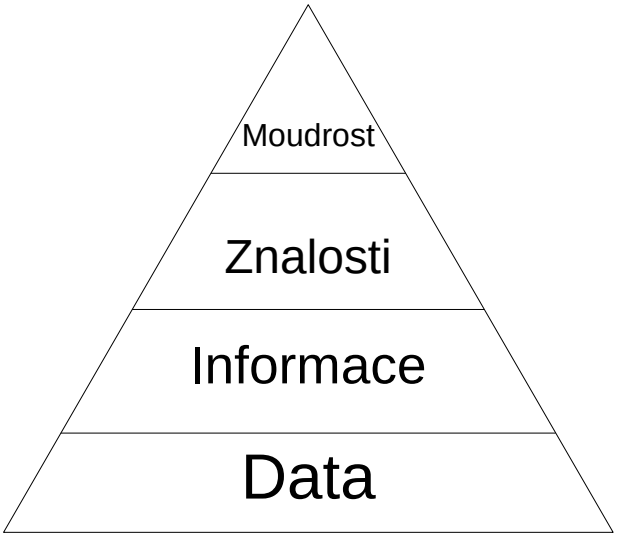
Luboslav Lacko ve své knize definuje Business Intelligence takto [15]:

Proces transformace dat na informace a převod těchto informací na poznatky prostřednictvím objevování nazýváme Business Intelligence.

Jinými slovy, účelem BI je konvertovat velké objemy dat na poznatky, které jsou potřebné pro koncové uživatele. Právě tyto poznatky jsou následně efektivně využívány při procesu rozhodování.

Pojem informace nepředstavuje pouze konkrétní záznam nebo množinu záznamů. Proces transformace dat na informace, informací na znalosti a budování „moudrosti“ na základě znalostí, lze přehledně zobrazit na hierarchické pyramidě informačních úrovní (*Obrázek č. 5*).

Základ pyramidy a prakticky všeho jsou data. Ta obsahují pouze jednoduchá fakta. Intuitivně však lze předpokládat, že někde uvnitř množiny dat jsou ukryty informace. Informace jsou odhaleny až v momentě, kdy jsou k datům přidány souvislosti. Následně vstupuje do procesu tvořivá inteligence a informace jsou transformovány na znalosti. Zobecněním znalostí vzniká „moudrost“ – schopnost přesného zhodnocení znalostí a jejich následné uplatnění v reálné praxi.



Obrázek č. 5: Hierarchie informačních úrovní

5.1 Přehled databázových systémů

V současné době existuje mnoho databázových/informačních systémů, které lze ve společnosti používat. Každý z nich pokrývá určitou oblast činnosti a rozhodování společnosti. Každý z nich je určen pro jinou skupinu uživatelů v rámci podnikové hierarchie.

OLTP (On-Line Transaction Processing)

Data jsou ukládaná do transakčních databází. Ty jsou určeny pro vykonávání velkého množství transakcí – bankovních, obchodních a podobně. Takové databáze jsou pak propojeny na nejrůznější IT systémy, jejichž cílem je automatizace každodenních činností, které jsou předmětem podnikání společnosti. Mezi tyto IT systémy se například řadí skladová evidence, mzdy, nákup a prodej, případně řízení a monitorování technologických procesů v reálném čase. Tyto systémy jsou v některých oblastech firemní informatiky téměř nepostradatelné. Kromě nesporných výhod, které vyplývají z jejich principu (například rychlost zpracování požadavku), jsou v mnoha společnostech preferovány také díky existenci velkého počtu specialistů. V praxi rovněž dochází k průniku systémů. Pokud transakční databázový systém s příslušnými aplikačními nadstavbami pokrývá většinu podnikových aktivit, nazývá se *ERP systém (Enterprise Resource Planning).* U těchto systémů může ke zdroji v jednom okamžiku přistupovat velké množství uživatelů. Někteří z nich data zapisují, jiní pouze čtou. Toto je jeden z úkolů, proč takovéto systémy nejsou vhodné pro vykonávání analýz.

Systémy MIS (Management Information Systems)

Jde o systémy pro podporu, řízení a rozhodování. Do těchto systémů vstupují data z transakčních systémů. MIS poskytují řídícím pracovníkům různé komplexní přehledy a sestavy, agregované podle různých hledisek, například časových, geografických, organizačních a jiných. Nevýhodou zde je poměrně velká režie. Požadavky na konkrétní sestavu se odesílají vývojářům MIS, kteří sestavu vytvoří a poskytnou ji manažerům až po určité době. Zpravidla se jedná o dny, týdny nebo dokonce i měsíce.

Systémy DSS (Decision-Support Systems)

Tyto systémy nabízí operativnější výsledky než MIS. Již z názvu je patrné, že tyto výsledky jsou určeny pro podporu rozhodování. Na rozdíl od systému MIS, které jsou nasazovány na operativní taktické úrovni, DSS už jsou na rozhraní taktického a strategického rozhodování. Mimo jiné nabízí řídícím pracovníkům výsledky poměrně složitých analýz, které mohou být právě důležité při taktických či strategických rozhodnutích společnosti.

Systémy EIS (Executive Information Systems)

Nejvyšší úroveň pak tvoří systémy EIS, které představují informační systémy pro vrcholové řízení. Mnohem častěji se však vyskytují pod termínem Business Intelligence. Účel BI byl popsán dříve, jde tedy o konverzi velkých objemů dat na poznatky, které jsou potřebné pro koncové uživatele. Tyto poznatky lze potom efektivně využívat například v procesu rozhodování.

5.2 Kvalita údajů pro analýzy

Společnosti používají pro svou činnost různé druhy ekonomického softwaru, například účetnictví, skladové hospodářství, evidence pohybu zboží a podobně. Při této činnosti se hromadí velké množství dat. Může se jednat o zcela bezcenná data, ale možná i velmi cenná. Zůstávají však nevyužita, protože jsou uchovávána ve formě, díky které jsou nedostupná pro potřeby získávání informací. To, že společnost vlastní data, ještě neznamená, že rovněž disponuje informacemi.

Pokud společnost prodává 1000 výrobků, prostřednictvím deseti kanálů celkově 100 odběratelům, tak již při tak malém počtu výrobků existuje milión kombinací uvedených položek. V momentě, kdy je požadováno sledovat obchodní život firmy v jednotlivých měsících, počet kombinací již dosahuje čísla dvanácti miliónů. V případě, že je potřeba sledovat ještě další ukazatele, počet kombinací a tím pádem ji složitost a rychlost výpočtu požadovaných informací rapidně roste. Řešením je použití tzv. multidimenzionálních databází, které jsou právě pro takovéto účely navrženy a optimalizovány. Podrobně se této problematice včetně všech jejich úskalí věnuje kapitola zaměřena na datové sklady.

5.3 Nevhodnost transakčních databází pro analýzy

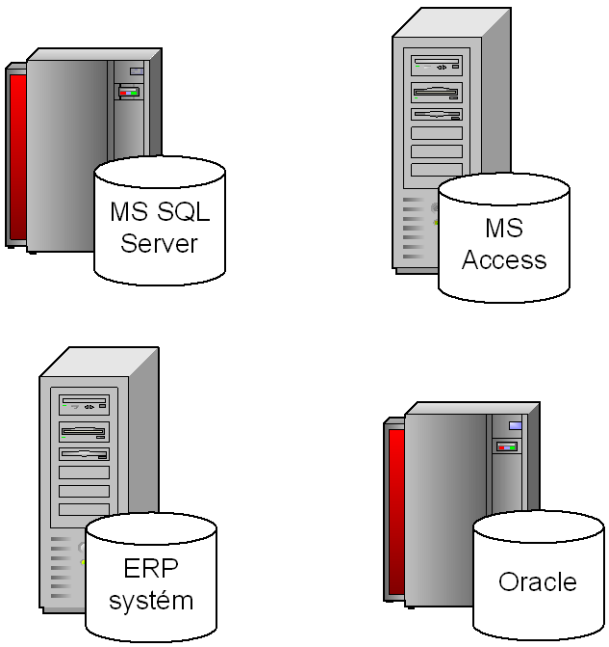
Transakční databáze označované i jako OLTP databáze jsou v podnikové praxi, ale i jinde používány pro ukládání operativních údajů. Data jsou uložena napříč tabulkami, souhrny, různé sestavy a podobně jsou získány pomocí agregačních funkcí některého z dotazovacích jazyků.

Z důvodu jednoduchého a rychlého dotazování, vyloučení redundance, jsou z OLTP databáze normalizované. To znamená, že vyhovují pravidlům normálních forem. Data v těchto databázích jsou tedy komplexně a vysoce strukturovaná – splňují pravidla třetí normální formy (3NF). Díky tomu tyto systémy dosahují vysokého výkonu při on-line transakcích (operativních dotazech) než při složitých analýzách, které jsou zpravidla velice náročné na výpočetní výkon. Tyto komplexní analýzy vyžadují jiné techniky návrhu databází. Například použití multidimenzionálních a hvězdicových schémat s tabulkami faktů a dimenzí.

Čili lze tvrdit, že OLTP databáze jsou optimalizovány pro obchodní transakce, ale nejsou vhodné pro získání informací pro podporu rozhodování, jednak z důvodů popsaných výše, ale také díky dalším vlastnostem, které budou zmíněny nyní.

5.3.1 Decentralizovanost systémů OLTP

Dalším problémem při použití systémů OLTP pro analýzy je skutečnost, že tyto systémy nemají integrovaný zdroj údajů ze všech systémů v rámci podniku tak, aby umožnily vytvářet komplexní analýzy. Potřebná data nebo data, která by měla sloužit jako podklady pro analýzy, jsou rozmístěna v různých heterogenních OLTP systémech napříč podnikem a musí se pokaždé pracně integrovat dříve, než je možné získat požadované informace (*Obrázek č. 6*). Díky tomuto složitému procesu, který by musel pokaždé předcházet, je časová náročnost případných analýz velice vysoká a to se nemusí jednat ani o nijak složité či komplexní analýzy. V některých případech se dokonce může stát, že se nepodaří konsolidovat veškerá data ze všech zdrojů, takže společnost nemůže získat celistvý obraz o stavu svého podnikání.



Obrázek č. 6: Decentralizace systémů OLTP

Nevhodnost OLTP systémů pro analýzy podtrhují i následující vlastnosti:

- **transakční systém neuchovává historická data**
  Historická data jsou potřebná pro komplexní analýzu či predikci. Ne vždy je však OLTP systém schopen tato data uchovávat. Jedním z důvodů může být nedostatečná disková kapacita databázového serveru určeného pro sběr dat.

- **nehomogenní struktura dat**
  Totožná data mohou být v různých systémech uložena v různých formátech a tvarech. Například problémy bývají s rodným číslem, kdy jednou může být uloženo jako desetimístný řetězec, jinde je po prvních šesti číslicích lomítko a poté další čtyři číslice. Podobné je to také u telefonního čísla, kdy v jednom systému může být uloženo v mezinárodním formátu, jinde naopak pouze jako devítimístný řetězec. Takovýchto problémových dat může být celá řada.

- **dlouhý čas přípravy dat**
  Zpravidla bývají data, která jsou potřebná jako vstupní pro nejrůznější analýzy uložena právě v nehomogenních zdrojích. Postupné připojení k nim není z technického hlediska ve většině případů problém, ovšem tato operace zabere mnoho času a námahy. Kromě tohoto technického hlediska existují také bariéry zapříčiněné lidským faktorem. Analytici totiž zpravidla neovládají dotazovací jazyk SQL, proto musí analýzy vypracovávat ve spolupráci s databázovými specialisty. Toto opět může trvat dlouho, navíc to neúměrně zvyšuje náklady.

Také díky těmto nedostatkům, kterými trpí klasické transakční databáze při jejich nasazení jako nástroje pro vytváření komplexních analýz, byl navržen systém **OLAP (On-line analytical processing)**, který je navržen tak, aby tyto nedostatky eliminoval. Jde o systém, který je používán pro vykonávání komplexních analýz, úzce souvisí s datovými sklady a podrobně se mu věnuje následující kapitola.

6 OLAP a datové sklady

Minulá kapitola popisovala nevhodnost klasických operativních databází pro potřeby vykonávání komplexních analýz, jejichž výsledky se používají jako podpora při strategických, taktických rozhodnutích. Z těchto nedostatků vyvstala potřeba nového systému, služeb a prostředků, které budou speciálně navrženy přesně pro tyto potřeby. Tyto potřeby pokrývá problematika OLAP (On-line analytical processing) a datové sklady, kdy tyto dva pojmy jsou velice úzce spjaty.

Vysvětlení základních principů, terminologie používané v této oblasti, ale i popis technik používaných při budování datového skladu s využitím OLAP včetně praktické ukázky na reálném příkladě je cílem této kapitoly.

6.1 Úvod do problematiky OLAP

Pojem OLAP definoval Dr. E. F. Codd, jeho cílem bylo popsání technologie, která by pro řízení podnikových dat více využívala osobních počítačů. Pro OLAP existuje mnoha definic, například [15]:
OLAP je volně definovaná řada principů, které poskytují dimenzionální rámec pro podporu rozhodování.

Pojem OLAP bývá často zaměňován s pojmem DSS (Decision Support Systems) – systém pro podporu rozhodování.

6.1.1 Dimenze a fakta

Pro vytvoření datového skladu – tzv. OLAP kostky je potřeba dvou typů dat – faktů a dimenzí. Tato data jsou uchovávána v samostatných databázových tabulkách.

Definice [15]**:** *Fakta jsou numerické měrné jednotky obchodování.*

Tabulka faktů bývá ve většině případů největší tabulka v databázi, obsahuje velký počet dat. Tabulky faktů mohou vytvářet různá schémata – například hvězdicové schéma nebo schéma sněhové vločky. V případě hvězdicového schématu je obsažena pouze jediná tabulka faktů. U jiných typů schémat může samozřejmě být tabulek faktů několik, také dimenze mohou vytvářet různé hierarchie. Pomocí prvotních fakt, například výrobní cena, prodejní cena lze dopočítávat další – odvozená fakta, která lze rovněž využít v datovém skladu. V tomto případě by se jednalo o fakt zisk z prodeje. Na faktech se v rámci datového skladu aplikují jednotlivé agregační funkce (například součet, průměr, minimum, maximum, počet) a to napříč jednotlivými dimenzemi.

Definice [15]**:** *Dimenze obsahují logicky nebo organizačně hierarchicky uspořádaná data. Jsou to vlastně textové popisy obchodování.*

Tabulky dimenzí jsou zpravidla menší než tabulky faktů, rovněž v mnoha případech obsahují konstantní data, případně data, která se příliš nemění. Dimenze lze chápat jako „vysvětlovací“ atributy pro jednotlivé měrné jednotky obchodování (fakta). Nejčastěji se používají časové, geografické a subjektové (zákazník, produkt) dimenze.

6.1.2 Multidimenzionální databázový model

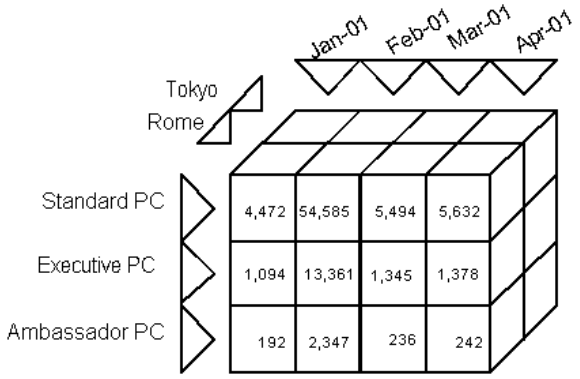
Řešením hlavních omezení relačních databází pro potřeby analýz je zavedení organizace údajů do multidimenzionálních struktur. Takto vytvořená databáze poté slouží jako základ pro uložení sumarizovaných a agregovaných údajů – informací. Do této multidimenzionální datové struktury se ukládají již očištěné a jinak upravené údaje. Proces čištění a úpravy dat je součástí problematiky budování datového skladu a je blíže rozebrán později. V tomto případě se na rozdíl od relačních databází používají téměř výhradně nenormalizované tabulky s redundantními a dopředu předpočítanými údaji. Základem datového skladu tedy je multidimenzionální databáze.

Ta má své výhody i nevýhody. Mezi výhody patří:

- rychlý a komplexní přístup k velkému objemu údajů,
- přístup k multidimenzionálním a relačním datovým strukturám,
- možnost komplexních analýz,
- silné schopnosti pro modelování a prognózy.

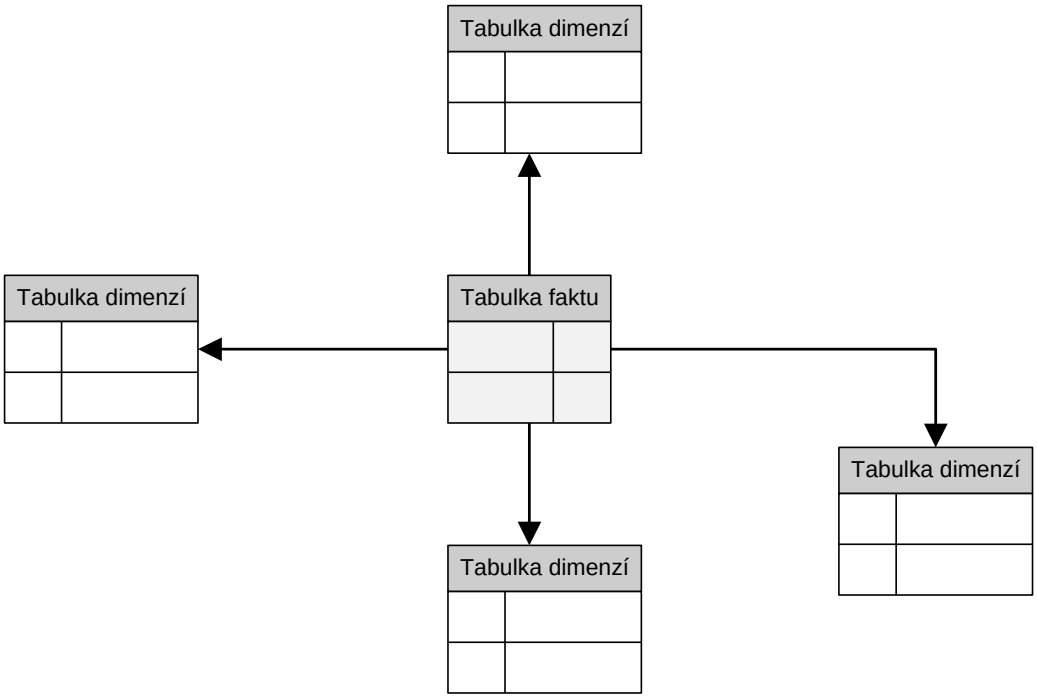
Naproti tomu nevýhodou může být například vyšší nároky na kapacitu úložiště.

Data, která jsou součástí datového skladu, jsou tedy uloženy v tabulkách dimenzí a faktů. Na základě těchto tabulek je pak vytvořena multidimenzionální datová struktura – **kostka**. Dimenze představují osy kostky, údaje (fakta) se pak nachází v průnicích jednotlivých dimenzí. Takto lze procházet celou kostku a analyzovat tak měrné jednotky například pouze za určité období nebo oblast. Lépe si lze celou situaci představit v prostoru na klasické trojrozměrné kostce (*Obrázek č. 7*). Ve skutečnosti však počet dimenzí není nijak omezen.



Obrázek č. 7: Ukázka OLAP kostky

Kostka se vytváří na základě dimenzionálního modelu, který má dané topologické uspořádání. Toto uspořádání se nazývá *schéma*. Dříve již byly zmíněny často se používající schémata. Nyní pouze pro doplnění příklad hvězdicového schématu (*Obrázek č. 8*).



Obrázek č. 8: Hvězdicové schéma

Hvězdicové schéma se skládá s tabulky faktů a obsahuje cizí klíče, které se vztahují k primárním klíčům v tabulkách dimenzí.

6.1.3 Úložiště multidimenzionálních dat

Existuje několik druhů multidimenzionálních databázových modelů [17]. Tato kapitola si klade za cíl je popsat a zhodnotit jejich výhody a nevýhody.

Multidimenzionální OLAP (MOLAP)
Pro tento model se data získávají z datového skladu nebo z operačních zdrojů. Mechanismus MOLAP následně data ukládá ve vlastních datových strukturách. Během tohoto procesu se spočte co možná největší počet předběžných výsledků. Čili v tomto úložišti jsou data uložena jako dopředu vypočítaná pole. Databáze je organizována tak, aby rychle získávala data z jednotlivých dimenzí. Výhodou je v tomto případě vysoký výkon vzhledem k dotazům uživatele, nevýhodou naopak redundance dat. Ty jsou totiž uloženy jak v relačních strukturách, tak v multidimenzionální databázi. V případě, kdy datový sklad obsahuje velký počet dimenzí, požadavky na úložný prostor mohou také výrazně stoupat.

Relační databázový OLAP (ROLAP)
Tento model získává data pro analýzy z relačního datového skladu. Data se po zpracování předkládají uživatelům jako multidimenzionální pohled.

Data i metadata se v úložišti ROLAP ukládají jako klasické záznamy v relační databázi. Pro generování výsledků podle požadavků uživatelů, používá OLAP server právě tato metadata, ze kterých dynamicky generuje SQL příkazy pro získání potřebných dat. Výhodou tohoto modelu je, že předchází redundanci, protože data jsou uložena pouze v relačních databázích. Nevýhodou naopak může být nižší výkon při obsluhování dotazů uživatelů a zcela jistě i vyšší režie než u modelu MOLAP.

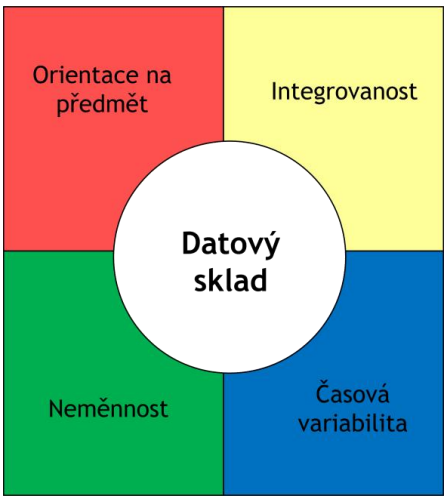
Hybridní OLAP (HOLAP)
Tento model je kombinací předešlých dvou, přičemž se snaží využívat pouze výhod obou úložišť a jejich nevýhody eliminovat. V tomto případě data zůstávají v relačních databázích a vypočítané hodnoty se ukládají do multidimenzionálních struktur.

6.2 Datový sklad

Datový sklad (DS) je prakticky strukturované úložiště dat. Nejznámější formální definice DS pochází od Billa Inmona [16] :

Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnlivých, historických dat použitých na získávání informací a podporu rozhodování (Obrázek č. 9). V datovém skladu jsou uložena atomická a sumární data.

Zdrojové údaje mohou být uloženy v různých operativních databázích a to dokonce v rozličných geografických lokalitách. Tyto údaje se v pravidelných intervalech sbírají, předzpracovávají a zavádějí do datového skladu, kde poté slouží pro podporu rozhodování.



Obrázek č. 9: Grafické vyjádření definice datového skladu

Uvedená definice je dostatečně výstižná, ale příliš stručná. Následuje rozbor jednotlivých pojmů, které se v ní vyskytují:

- **subjektová orientace:** Data jsou v datovém skladu ukládána spíše podle předmětu zájmu než podle způsobu jejich použití. Při orientaci na subjekt jsou data v datovém skladu kategorizována podle subjektu, který může například být: zákazník, výrobek, prodejna. Naproti tomu ukládání dat podle způsobu jejich použití znamená, že data jsou uložena tak, jak jsou reflektována v běžné praxi – například objednávky, personalistika a podobně.

- **integrovanost:** Data v datovém skladu musí být jednotná a integrovaná. To znamená, že údaje týkající se konkrétního subjektu se do DS ukládají jen jednou – dimenze musí obsahovat neredundantní data. Z toho důvodu je potřeba zavést jednotnou terminologii, sjednotit měrné jednotky a formáty jednotlivých atributů. Data totiž do DS přichází z nekonzistentního a neintegrovaného prostředí. Příprava, čištění a úpravy dat jsou součástí etapy nazvané ETL (Extract, Transformation, Loading). Této etapě je věnována samostatná kapitola práce.

- **časová variabilita:** Data se do datového skladu ukládají jako série obrazů, z nichž každý představuje určitý časový úsek. Na rozdíl od operativních databází, kde jsou data platná v okamžiku přístupu, v datových skladech jsou platná po určitý časový úsek. V operativních databázích se data ukládají za menší časové období – dny, hodiny. V datovém skladu naopak jsou uložena data za delší časové období – měsíce, roky. Datum v tomto případě může vytvářet různé hierarchie (například rok-kvartál-měsíc-týden a jiné). Díky multidimenzionálnímu datovému modelu lze data analyzovat napříč časem.

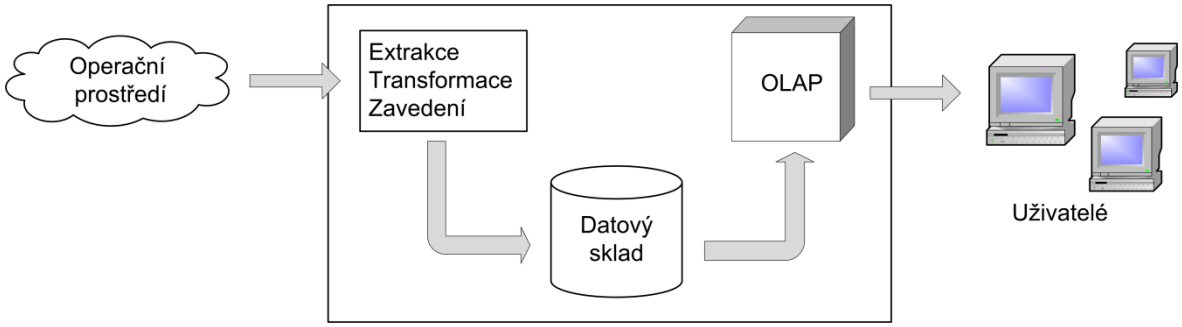
- **neměnnost:** V klasických operativních databázích jsou údaje vkládány, modifikovány a mazány. Běžná praxe v oblasti datových skladů je však jiná. Zde se údaje nemodifikují a zpravidla ani nemažou. Pouze se v pravidelných intervalech přidávají nové údaje. Z tohoto zjištění vyplývá, že jsou přípustné pouze dva typy operací. Nahrání údajů do DS a přístup k nim. Modifikace dat není přípustná.

Většinu rozdílů mezi operativními databázemi a datovými sklady shrnuje *tabulka č. 1* [16]:

Vlastnost	Operativní databáze	Datový sklad
Čas odezvy	Zlomky sekund až sekundy	Sekundy až hodiny
Operace	Kompletní manipulace s daty	Primární je čtení
Stáří dat	30-60 dní	Série snímků za časové období
Organizace dat	Podle způsobu použití	Podle předmětu, času
Zdroje dat	Operační, interní	Operační, interní, externí
Činnosti	Procesy	Analýza

Tabulka č. 1: Porovnání operativních databází a datových skladů

Datový sklad tedy slouží pro vykonávání nejrůznějších analýz, které jsou následně využívány pro potřeby rozhodování manažerů či obchodníků. Nástroje, které jsou potřebné k provozu a vybudování datového skladu, však vyžadují značnou investici, proto jsou tato řešení využívána zejména většími společnostmi, jako jsou banky pojišťovny či velké obchodní řetězce.
K údajům v datovém skladu mohou mít prostřednictvím webu přístup všichni zainteresovaní uživatelé a dokonce i obchodní partneři, pokud to společnost požaduje. Následující schéma (*Obrázek č. 10*) představuje zevrubný pohled na problematiku datových skladů a OLAP analýzy. Je potřeba získat údaje z operativního prostředí, vhodně je upravit, transformovat, následně zavést do datového skladu, kde se nad těmito daty vykonají analýzy a jejich výsledky se zpřístupní uživatelům.



Obrázek č. 10: Princip činnosti datového skladu

6.3 Metody budování datového skladu

Jedním z nejdůležitějších kroků při budování datového skladu je rozhodnutí o tom, jakou metodu použít. Je nutné brát ohled jednak na organizační strukturu společnosti, ale také na dostupnost a kvalitu požadovaných informací. Rovněž je vhodné předvídat problémy, které se mohou během budování datového skladu objevit. Cílem této kapitoly je tedy popsat nejznámější a nejčastěji používané metody budování datového skladu.

6.3.1 Metoda „velkého třesku“

Některé společnosti se domnívají, že je možné datový sklad vybudovat v rámci jediného projektu. Vývoj datového skladu je však náročný problém a s největší pravděpodobností se ho nepodaří vyřešit najednou a v rozumném čase. Toto je považováno za největší slabinu, jelikož pokud už by se datový sklad podařil pomocí této metody vybudovat, tak během té doby již může dojít ke změně technologie či požadavků uživatelů. Tato metoda se skládá ze tří etap:

- Analýza požadavků podniku
- Vytvoření podnikového datového skladu
- Vytvoření přístupu buď přímo, nebo prostřednictvím datových trhů

Metoda velkého třesku má pouze jedinou výhodu a to tu, že celý projekt lze vypracovat ještě před začátkem jeho samotné realizace. Ovšem není to výhoda v pravém slova smyslu – budování datového skladu je dynamický proces, dá se proto s velkou pravděpodobností předpokládat, že se minimálně během jeho realizace změní požadavky uživatelů.

Zcela jistě u této metody převažují nevýhody, které lze považovat za závažné:

- Velké riziko změny požadavků
- Velice dlouhá doba než jsou známy první výsledky – dlouhá doba, než je společnost schopna získat obchodní zisk plynoucí z nasazení datového skladu

6.3.2 Přírůstková metoda

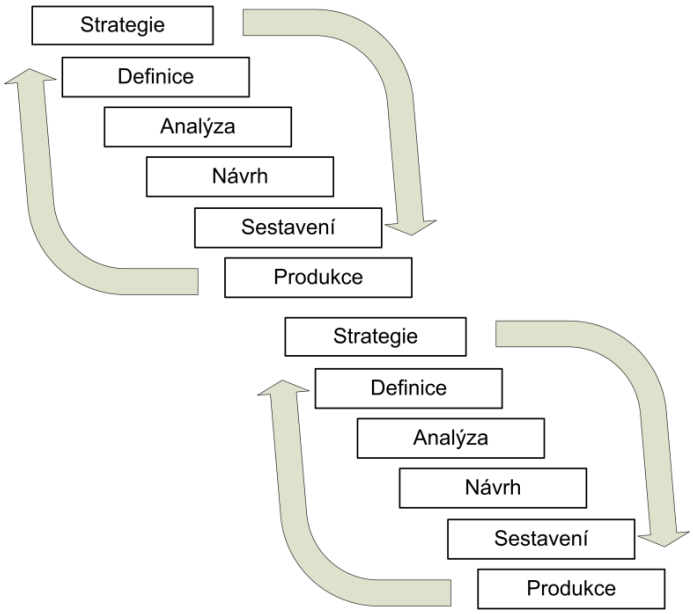
Ekvivalentní název metody je evoluční. Princip spočívá v budování datového skladu po jednotlivých etapách. Čili místo vybudování celého DS najednou postupně přibývají dílčí přírůstky, které samozřejmě zapadají do celkové architektury datového skladu. *Zde si lze všimnout analogie s iterativním vývojem softwaru.*

Budování datového skladu začíná tak, že se nejdřív vybuduje několik málo předmětných oblastí, které se implementuje např. jako datový trh. Ten je pak poskytnut koncovým uživatelům. Tím se částečně uspokojí hlad managementu po výsledcích, zkrátí se návratnost investice, jelikož první subsystémy DS začnou fungovat ne dlouho po spuštění celého projektu. Výběr předmětných oblastí, které budou předně analyzovány a navrženy, by měl být konzultován s koncovými uživateli a měl by reflektovat jejich požadavky. Tento proces se opakuje tak dlouho, dokud není vybudován celý datový sklad.

U této metody na rozdíl od předchozí převažují výhody nad nevýhodami. Mezi pozitiva patří:

- Přírůstkové budování DS zachovává kontinuitu budovaného projektu s požadavky a potřebami uživatelů.
- Umožňuje implementovat rozšiřitelnou architekturu
- Je zaručen rychlejší zisk a tedy i návratnost investice

Jde tedy o iterativní proces, který obsahuje pouze dvě iterace. Schematicky lze přírůstkovou metodu znázornit takto (*Obrázek č. 11*):



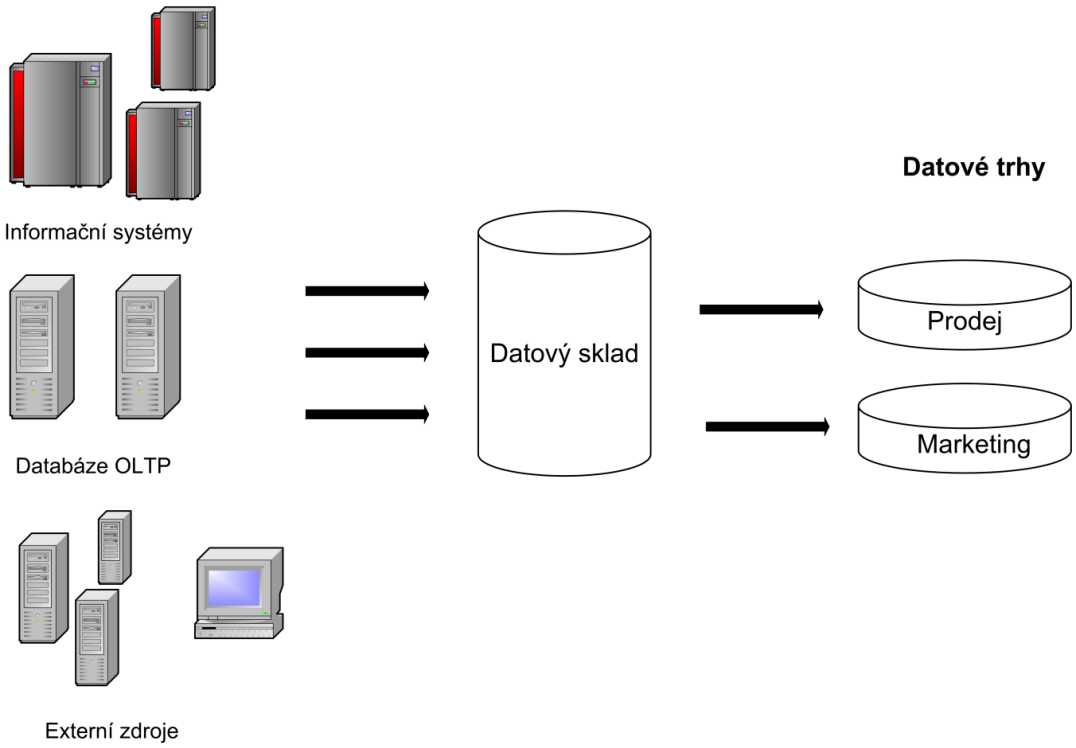
Obrázek č. 11: Schéma přírůstkové metody

Přírůstková metoda směrem „shora dolů“

Při budování DS pomocí přírůstkové metody lze volit jednou ze dvou variant. Varianta shora dolů je jedna z nich. Nejdříve se vytvoří konceptuální model datového skladu na základě požadavků uživatelů, přičemž se také stanoví předmětné hierarchie. Poté jsou vytvořeny konceptuální modely jednotlivých předmětných oblastí. Jinými slovy řečeno, postupně se vytvářejí datové trhy jednotlivých předmětných oblastí v rámci struktury datového skladu.

Tato metoda nabízí rychlou implementaci jednotlivých datových trhů a s tím související rychlejší návratnost investic. Oproti metodě velkého třesku tato přírůstková metoda shora dolů není tak náročná na analýzu. Mezi nevýhody patří zvýšené vstupní náklady dříve, než je možné předvídat návratnost investice.

Konceptuální náhled na tuto metodu nabízí následující schéma (*Obrázek č. 12*):

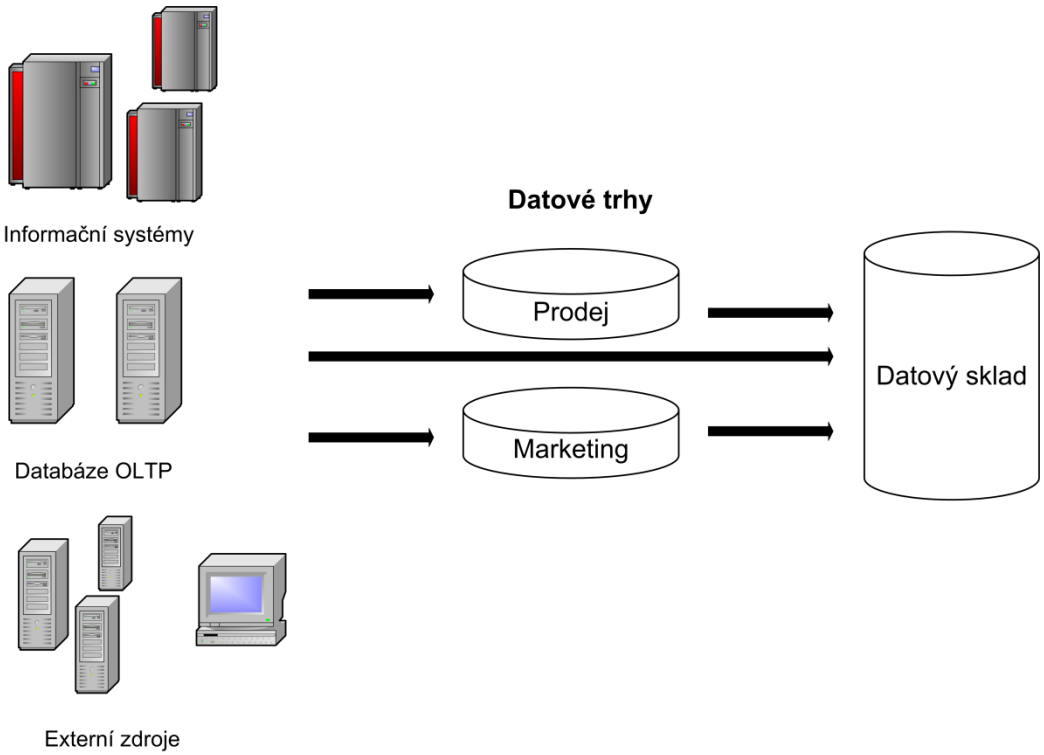


Obrázek č. 12: Schéma přírůstkové metody shora dolů

Přírůstková metoda směrem „zdola nahoru“

Tato metoda je velice podobná metodě shora dolů s tím rozdílem, že v tomto případě mají větší prioritu údaje před obchodním ziskem. To znamená, že se nejdříve budují datové trhy předmětných oblastí v rámci struktury datového skladu (*Obrázek č. 13*).

U této metody hraje podstatnou roli IT oddělení dané společnosti. I když se konceptuální model odvíjí od zdrojových systémů, bývá v některých případech celková rozšiřitelnost značně problematická. IT oddělení totiž není v mnoha společnostech považováno za „důležité“, co se týče strategie a marketingu, proto se o mnohých připravovaných změnách a záměrech dozvídá zpravidla jako poslední. Z toho důvodu mohou navrhnout a dokonce i realizovat něco, co je vzhledem ke strategickým plánům společnosti již neaktuální.



Obrázek č. 13: Schéma přírůstkové metody zdola nahoru

Ze schématu přírůstkové metody (*Obrázek č. 11*) vyplývá, že metoda se skládá z následujících kroků:

- Strategie
- Definice
- Analýza
- Návrh
- Sestavení
- Produkce

Strategie

Úloha této fáze je pouze jedna. Dalo by se říci, že jde o klíčovou věc při budování datového skladu. Je potřeba definovat cíle. Na jedné straně je to cíl podnikání, na druhé účel řešení datového skladu. Kvůli těmto podstatným rozhodnutím, by tato fáze měla být plně v rukách vrcholového managementu společnosti. Cíle z pohledu podnikové strategie mohou být krátkodobé a dlouhodobé. Dlouhodobé cíle projektu datového skladu zahrnuje jeho vybudování, ale také definuje jeho správu a rovněž nejsou opomenuty další významné kroky, jako je vypracování a správa dokumentace datového skladu, školení jeho uživatelů. Tato fáze zahrnuje i základní definici architektury podnikového DS.

Definice

Tato fáze se snaží definovat rozsah a cíl přírůstkového vývoje. Bude vytvořen počáteční přírůstek, konceptuální modely, zdokumentují se zdroje dat a vymezí se rozsah kvality těchto údajů.

Rovněž tato fáze zahrnuje návrh architektury DS a technických prostředků. Dále během této fáze by měly být dostatečně pochopeny struktury operačních a externích zdrojů dat. Na závěr se stanoví krátkodobé a dlouhodobé obchodí cíle, pro jejichž podporu je datový sklad budován.

Analýza
Cílem fáze analýzy je zaměřit se na informace o uživatelích, získávání dat a také na požadavky spjaté s přístupem k datům. Dále se vytvářejí relační a multidimenzionální datové modely pro datový sklad. Během této fáze by měl být také dokončen výběr nástroje pro všechny komponenty datového skladu. Rovněž by se v této fázi měly řešit problémy s kvalitou dat.

Návrh
Cílem fáze návrhu je transformovat požadavky získané během fáze analýzy do detailních podmínek návrhu a dokončit instalaci technické architektury.

Sestavení
Tato fáze si klade za cíl vytvořit a otestovat navržené databázové struktury, moduly pro získávání dat, moduly správy datového skladu, moduly metadat, moduly přístupu k datům, sestavy a dotazy.

Produkce
Závěrečná fáze představuje fázi přechodu do produkce. V této fázi dochází k instalaci datového skladu, ten se začíná používat. S tím souvisí i nutnost řešit údržbu a řízení růstu datového skladu.

Jak je patrné z textu a schémat výše, problematika budování datového skladu vychází ze stejných principů, jako iterativní vývoj jakéhokoliv jiného softwarového produktu podle metodologie RUP. Činnosti náležící jednotlivým fázím jsou prakticky totožné, v tomto případě pouze modifikovány pro potřeby budování datového skladu. Dalším rozdílem může být počet iterací, kde v případě datových skladů se počítá pouze s dvěma iteracemi [16].

6.4 Etapa ETL

Nasazení technologií business intelligence prakticky nikdy nezačíná na zelené louce. Pokud společnost se snaží již nějak řešit podporu pro rozhodování, obvykle se využívá údajů pocházejících z transakčních systémů OLTP. V lepším případě jsou tyto údaje zpracovány do sestav například pomocí softwaru Microsoft Office.

Údaje pro business intelligence, a tedy i datový sklad, pochází z různých nehomogenních prostředí. Může se jednat o údaje ze souborových databází jako je například MS Access nebo se může jednat o údaje z databází spravovaných některým z databázových serverů (například Oracle, MySQL, Microsoft SQL Server a podobně). Jak již bylo zmíněno dříve, *příprava a zavedení dat je důležitou součástí každého řešení datového skladu*. Veškeré údaje je potřeba z operačních prostředí vyextrahovat, očistit, upravit a teprve poté zavést do datového skladu. Tyto úkony jsou právě cílem etapy ETL (Extract, Transformation, Loading). Cílem této kapitoly je popsat tuto důležitou etapu budování datového skladu, popsat její jednotlivé fáze a zmínit také problémy, které během této etapy mohou nastat.

Jednotlivé fáze etapy ETL lze do češtiny přeložit jako – *extrakce, transformace, zavedení*. Jinými slovy jde o přípravu údajů. V některých odborných publikacích se lze rovněž setkat s termínem *datová pumpa*. Nástroje a postupy související s etapou ETL jsou součástí každého projektu datového skladu. Celý tento proces je komplexní a ve většině případů také časově náročný. Docela běžně se stává, že etapa ETL zabere polovinu času z celkového času projektu budování DS. S tím souvisí i největší vynaložené úsilí a velká část potřebných nákladů. Při bližším pohledu na jednotlivé fáze etapy:

- **Extrakce** – výběr dat prostřednictvím různých metod,
- **Transformace** – ověření, čištění, integrování a časové označení dat,
- **Zavedení** – přemístění dat do datového skladu

lze zjistit, že hlavím cílem etapy ETL je centralizace údajů. Tím je myšleno jejich shromáždění z několika nehomogenních zdrojů a naplnění datového skladu v požadovaném čase. Údaje se v této etapě nejen přenášejí, ale i zpracovávají – například indexují, sumarizují, zjišťují se případné změny struktur zdrojových dat, které jsou potřebné pro datový sklad, dále se udržují metadata, což v tomto případě znamená předpisy a definice pro přenos a zpracování údajů.

Etapa ETL však nekončí pouze prvotním naplnění datového skladu, ten se v pravidelných intervalech neustále plní novými daty. Údaje, které se zavádějí do datového skladu, by měly být kvalitní, přesné a aktuální, což znamená především užitečné pro uživatele.

6.4.1 Extrakce

Údaje, které mají být obsaženy v datovém skladu, jsou jednak umístěny v nehomogenních operačních prostředních, hardwarových platformách (PC, mainframe, Mac), operačních systémech (Windows, Linux, Solaris, Sun), databázových systémech (MySQL, MS SQL, Oracle, IBM DB2), archivních systémech, podnikových systémech (SAP), a co je horší, zpravidla se vyskytují v různých formátech.

Samostatnou problematiku představují archivní data, která obsahují historické údaje. Hlavní rozdíl mezi datovým skladem a archivem je ten, že v datovém skladu se data pravidelně aktualizují. Údaje z archivů jsou však důležitým zdrojem historických dat a zpravidla bývají použita při prvním naplnění datového skladu.

V některých případech je žádoucí v datovém skladu pracovat i s externími údaji. Ty lze získat mnoha způsoby, například analýzou konkurenčního prostředí, zakoupením údajů o zákaznících, nebo je rovněž možné stáhnout volně dostupné informace na Internetu. Z povahy dat zde vyplývá, že u tohoto typu dat není možné periodicky data odebírat jako je tomu u interních zdrojů. Proto externí údaje vyžadují trvalé monitorování, aby bylo možné určit, kdy jsou dostupné a lze je zavést do DS.

6.4.2 Transformace

Data, která tvoří datový sklad, musí být kvalitní. Pokud DS obsahuje nekvalitní údaje, tak to snižuje důvěru v takovéto řešení a to se vcelku oprávněně přestane používat. Nekvalitní údaje se vyskytují již ve zdrojových datech.

Pokud jsou takováto nekvalitní či nepřesná data přenesena do datového skladu může to vést k chybným nebo nepřesným sestavám, z čehož následně můžou plynout chybná obchodní rozhodnutí.

Úkolem této fáze tedy je čištění zdrojových dat a snaha eliminovat data nekvalitní. Během této fáze může také docházet k nejrůznějším transformacím, což může představovat odvozování nových informací na základě těch dostupných. Například z data prodeje 2. 2. 2010 lze odvodit nové informace, jako jsou název měsíce (Únor), den v týdnu (úterý), číslo týdne, kvartál a podobně. Tedy mimo jiné je důležité rozložit zdrojová data na atomické atributy. Čištění údajů může být velmi náročné a nákladné. V některých případech dokonce ani nemá smysl čistit údaje s vysokými náklady, když je přínos pro podnikání zanedbatelný.

Transformace jako taková je tedy soubor úloh a úkonů, které vedou ke zvýšení kvality dat a hlavně k odstranění anomálií. Vybudování OLTP systému společnosti totiž zpravidla trvá několik let, neobvyklé případy nejsou ani ty, kdy společnost například používá již nějaký OLTP systém 20 let. Během té doby se zcela jistě obměňovaly verze softwarů, vývojová prostředí a platformy, na kterých se software vyvíjí. Mění se operační systémy a zpravidla také lidé, kteří se systémem pracují. Všechny tyto aspekty zásadně ovlivňují kvalitu dat a cílem této etapy, je tato data sjednotit a očistit.

Tato problematika však sebou nese celou řadu problémů, které mohou nastat. Nejčastěji se vyskytují problémy z těchto okruhů:

- Nejednoznačnost dat
- Chybějící hodnoty
- Duplicitní záznamy
- Nejednotnost názvů pojmů a objektů
- Nejednotnost peněžních měn
- Nejednotnost formátů čísel a textových řetězců
- Problémy s referenční integritou
- Chybějící datum

Nejednoznačnost dat

Například údaje o pohlaví zaměstnanců mohou být uložena různým způsobem *(muž, žena, M, F)*

Chybějící hodnoty a duplicitní záznamy

Problémy také činí chybějící hodnoty (NULL), případě duplicitní záznamy. Ty navíc mohou být otevřené nebo skryté. Větší problém však představují chybějící hodnoty, duplicitní záznamy se dají vždy odstranit, může to však být velice časově náročné. Pro odstranění chybějících hodnot existuje několik strategií. Pokud se jedná o malý počet údajů, je možné to ignorovat. Další možnosti je doplnění hodnot z jiných zdrojů (pokud je to možné) nebo rovněž lze chybějící hodnoty označit nějakou výchozí hodnotou a pracovat poté s takto označenou položkou, či ji ponechat označenou v OLTP systému a zapracovat později.

Nejednotnost názvů pojmů a objektů

V případě, že jsou slučovány údaje o stejném objektu z více zdrojů, kde každý má zavedené své pojmenovávání a terminologii, je nutné názvosloví a terminologii sjednotit.

Nejednotnost peněžních měn
Hodnoty měny mohou být také zdrojem problémů a nesrovnalostí, například hodnota 100,50 může znamenat úplně něco jiného v dánských korunách než ve forintech. Opět i zde je nutné sjednotit formáty měn a způsob jejich zápisu.

Nejednotnost formátů čísel a textových řetězců
V relačních databázích se údaje ukládají v různých druzích formátů. Číselné údaje totiž mohou být uloženy buď jako numerické hodnoty nebo jako řetězec znaků. V případě numerického datového typu se číslo ukládá jako numerická hodnota, do řetězcového datového typu se ukládá jako posloupnost čísel a případně i jiných znaků (*lomítko, desetinná čárka atd*.). Problémy potom mohou činit rodná čísla, kdy někdy jsou uložena jako desetimístné číslo, jindy zase po šesti číslech následuje lomítko a teprve potom závěrečná čtveřice čísel. Dalším problémovým atributem může být hodnota poštovního směrovacího čísla (PSČ), kdy jednou je tato hodnota uložena jako pěti místné číslo (11000), v jiném případě může být uloženo s mezerou (110 00) nebo s údajem, že se jedná o PSČ (PSČ 110 00). Kombinací existuje nesčetně mnoho a opět je nutné zvolit jeden způsob zápisu a ostatní položky tomu přizpůsobit.

Problémy s referenční integritou
Kromě samotných hodnot mohou být v datech skryty i různé vztahy. Například hierarchická struktura firmy nebo hierarchická struktura zaměstnanců. Jde však o dynamické údaje, které v praxi mohou často měnit. Pokud dojde ke zrušení nějakého oddělení v rámci firmy, může to zkreslit údaje, a tím pádem ovlivnit jejich kvalitu. Proto je nutné věnovat pozornost podobným situacím a umět na ně reagovat.

Chybějící datum
Čas hraje v datových skladech významnou roli. Skoro každá analytická databáze má časovou dimenzi. Může nastat problém, že v některých zdrojových databázích nemusí hrát čas významnou roli, v jiných je však čas významnou veličinou. Časový údaj musí být v datech přítomen před jejich zavedením do datového skladu. Pokud tomu tak není, musí se určit a přidat při zavádění dat.

6.4.3 Přenos

Poslední fází etapy ETL je přenos údajů z paměti zdrojových dat nebo dočasných úložišť do datového skladu. Samotný přenos spočívá v přesunu údajů a jejich uložení do databázových tabulek datového skladu. Přenos by měl být plánovaný a automatizovaný. Zejména při prvním plnění datového skladu může jít o obrovské množství dat. Následně již se údaje do datového skladu zavádějí v pravidelných intervalech a po menších objemech. Přenáší se pouze takové objemy dat, které za dané období v OLTP databázích vznikne. Navíc nepřenáší se všechny údaje, stále pouze ty, které jsou relevantní z hlediska využití v datovém skladu.

7 Vlastní řešení datového skladu

V rámci této kapitoly bude popsáno vlastní řešení datového skladu, a to od prvotní analýzy zdrojových dat – rozhodnutí o tom, které atributy jsou relevantní z hlediska využití v datovém skladu, přes návrh databázových struktur pro datový sklad. Kapitola rovněž obsahuje popis a způsob řešení etapy ETL. Následně jsou uvedeny výsledky OLAP analýzy a možnosti prezentace těchto výsledků koncovým uživatelům. Během následujícího textu je také popsána architektura, na které byl datový sklad budován. Cílem kapitoly mimo jiné je popsat vlastní řešení datového skladu a zhodnotit výsledky, které lze získat prostřednictvím OLAP analýzy.

7.1 Zdrojová data

Jako zdroj dat pro budovaný datový sklad byly použity data z aukčního systému. Konkrétně se jedná o údaje z prodeje zboží z oblasti numismatiky. Na úvod základní údaje o datech (*Tabulka č. 2*):

počet záznamů	493 608
počet atributů	36
počet tabulek	2

Tabulka č. 2: Základní údaje o datech

Jak je zmíněno výše (*Tabulka č. 2*), zdrojovou databázi tvoří dvě tabulky. První z nich má název *auction_price_ripper* a obsahuje údaje o jednotlivých aukcích. Celkový počet záznamů v této tabulce je 493 263.

Atributy:
(auction_id,auction_price, auction_name, auction_starting_price, auction_is_buy_now, auction_buy_now_price, auction_bid_count, auction_starting_time, auction_ending_time, auction_time_left, auction_category_id, auction_country, auction_seller_id, auction_seller_login, auction_bidder_id, auction_bidder_login, auction_description, auction_status, coin_label, coin_bid, coin_bid_before_end, auction_pictures, auction_pictures_small, coin_label_note)

Název druhé tabulky je *auction_price_ripper_bidder*. V této tabulce se nacházejí údaje o jednotlivých příhozech k aukcím. Celkový počet záznamů v této tabulce je 345.

Atributy:
(auction_id, user_id, user_login, user_rating, user_status, bidding_item_quantity, bidding_item_amount, date_of_bidding, bid_status, date_of_canceled, canceled_reason, canceled_bid_status)

7.1.1 Popis zdrojových dat

Tento oddíl textu blíže popisuje jednotlivé atributy zdrojové databáze a jejich vztah k nově vznikajícímu datovému skladu (*Tabulka č. 3*):

Název atributu	Popis	Ukázka hodnoty	Vztah k DS
auction_.id	Primární klíč tabulky, Identifikace jednotlivých aukcí	385152085	dimenze
auction_price	Prodejní cena	345.00	fakt
auction_name	Název aukce. Text, který je zobrazen uživateli ve výpisu aukcí.	Ševčínský důl PROOF !!! jen 5000ks !!!	dimenze
auction_starting_price	Vyvolávací cena položky.	99.00	fakt
auction_si_buy_now	Logický atribut, zda je aukce typu kup teď.	0	dimenze
auction_buy_now_price	kup teď cena	1000.00	fakt
auction_bid_count	počet příhozů u aukce	12	fakt
auction_starting_time	datum začátek aukce	1251564631	datumová dimenze
auction_ending_time	datum konec aukce	1251565367	
auction_time_left	čas do konce aukce	1251564631	nepoužito
auction_category_id	identifikátor kategorie	4675	dimenze
auction_country	identifikátor země	56	nepoužito
auction_seller_id	identifikátor prodejce	9783086	dimenze
auction_seller_login	login prodejce	Pepa_555	dimenze
auction_bidder_id	identifikátor zákazníka	6783433	dimenze
auction_bidder_login	login zákazníka	Franta332	dimenze
auction_description	popis zboží	Souvislý text	nepoužito
auction_status	status aukce – prodáno	sold	dimenze
coin_label	Tyto atributy prakticky ve všech záznamech obsahují NULL položky. V případě obrázků jsou někdy uvedeny dočasné adresy použitých obrázků, které se ale po určité době mažou. Z hlediska DS jde o nezajímavé atributy		nepoužito
coin_bid			
coin_bid_before_end			
auction_pictures			
auction_pictures_small			

coin_label_note	Stejný případ jako je zmíněno výše		nepoužito
user_id	identifikátor přihazujícího	6783433	dimenze
user_login	login přihazujícího	Pepa_555	dimenze
user_rating	hodnocení uživatele	155	dimenze
user_status	status uživatele – aktivní, neaktivní	0	nepoužito
bidding_item_quantity	počet položek, na které uživatel přihazuje	1	nepoužito
date_of_bidding	datum příhozu	1251564631	datumová dimenze
bid_status	status příhozu – např. zrušen, schválen	0	dimenze
date_of_canceled	datum zrušení příhozu	1251564631	nepoužito
canceled_reason	důvod zrušení příhozu	NULL	nepoužito
canceled_bid_status	status zrušeného příhozu	0	nepoužito

Tabulka č 3: popis zdrojových dat

Předchozí tabulka (*Tabulka č. 3*) tedy nabízí bližší pohled na zdrojová data. Mimo jiné z této tabulky lze vyčíst vztah zdrojových atributů vzhledem k nově vznikajícímu datovému skladu. Pro větší přehlednost nyní následuje přehled rozdělení atributů na dimenze a fakta:

Dimenzionální atributy:
auction_id, auction_name, auction_is_buy_now, auction_status, auction_seller_id, auction_category_id, auction_seller_login, auction_bidder_id, auction_bidder_login, user_id, user_login, user_rating, bid_status

Fakta:
auction_price, auction_starting_price, auction_buy_now, auction_bid_count

Nepoužité atributy:
Samozřejmě ne všechny atributy zdrojové databáze má smysl evidovat v datovém skladu. Následující tabulka (*Tabulka č. 4*) obsahuje přehled nepoužitých atributů včetně důvodu, proč daný atribut nebyl použit.

Název atributu	Důvod nezařazení
Auction_country	Jde o konstantní hodnotu. Ve všech záznamech figuruje stejná země. Pro potřeby datového skladu se tedy jedná o nezajímavý atribut.
Auction_descripton	Pro potřeby datového skladu se jedná o nezajímavý atribut. Pouze detailněji popisuje určitou dimenzi.
Coin_label, coin_bid, coin_bid_before_end, coin_label_note	Záznamy jsou prakticky u všech záznamů nulové.

Auction_pictures, auction_pictures_small	Hodnoty těchto atributů obsahují URL adresu k obrázku, respektive miniatuře k jednotlivým aukcím. Provozovatel aukce obrázky po určité době po skončení aukce maže – hodnoty se stávají neaktuálními. Pro potřeby DS opět nezajímavé atributy.
User_status	Konstantní hodnota. Ve všech případech se jedná o aktivní uživatele. Neaktivní uživatelé se nemohou účastnit aukcí. Z hlediska potřeb DS jde opět o neužitečnou informaci.
Bidding_item_quantity	Konstantní hodnota. Pouze ve dvou případech ze všech uživatel přihazuje na jiný počet předmětů než je jeden.
Date_of_canceled, canceled_reason, canceled_bid_status	Nedostatek dat. Ze všech případů příhozů na aukci, byl příhoz zrušen pouze jednou.

Tabulka č. 4: Nepoužité atributy

Doposud nebyly zmíněny atributy datumové. Tyto atributy vytvoří novou datumovou dimenzi. Rozdělení atributů do dimenzionálních tabulek a tabulek faktů následuje nyní.

7.1.2 Rozdělení atributů do dimenzionálních tabulek

- *dim_datum*
  (datum_id, datum_ansi, datum_full, rok, kvartál, měsíc, měsíc_název, týden_číslo, den_v_roce, den_v_týdnu, den_název)
  - Datumové údaje budou sdruženy v této dimenzionální tabulce. Vytvořit lze i několi hierarchií například den-měsíc-kvartál. Problematika naplnění této, ale i ostatních tabulek datového skladu je detailně popsána v části věnující se etapě ETL.

- *dim_aukce* (aukce_id, název, je_kup_teď, aukce_status)
  - informace o jednotlivých aukcích

- *dim_prodejce* (prodejce_id, prodejce_login)
  - informace o prodejcích

- *dim_zákazník* (zákazník_id, zákazník_login)
  - informace o kupujících

- *dim_přihazující* (přihazující_id, přihazující_login, hodnocení)
  - informace o přihazujících

Poznámka:
Zdrojová databáze obsahuje ještě atribut *auction_category_id*. Tento atribut představuje kategorii, do které je příslušné zboží zařazeno, protože v celé zdrojové databázi se vyskytlo pouze 37 různých kategorií a podle těchto identifikačních čísel se na webu provozovatele daly dohledat jména kategorií, byla vytvořena ještě nová dimenzionální tabulka:

- *dim_kategorie* (kategorie_id, obor, oblast, název_kategorie)
  - Tato tabulka tedy obsahuje jednotlivé kategorie, navíc lze uvnitř nalézt hierarchii. Obor představuje v hierarchické úrovni nejvyšší úroveň (např. Numismatika). Atribut oblast je další úrovní hierarchie a označuje například to, odkud daná mince pochází (Evropa, Česko atd.). Nejnižší úrovní je pak samotný název dané kategorie.

7.1.3 Tabulky faktů

- *fakt_statistika*
  (fakt_id, prodejní_cena, vyvolávací_cena, kup_teď_cena, počet_příhozu, nárůst_ceny)

Poznámka:
Nárůst_ceny je nový atribut odvozený z již dostupných atributů. Jde o číselnou hodnotu v procentech, která vyjadřuje procentuální nárůst ceny. Čili o kolik procent se zvýšila cena zboží oproti vyvolávací ceně. Hodnota je spočítána pomocí následujícího vzorce:

(Prodejní_cena – vyvolávací_cena) * 100 / vyvolávací cena = nárůst ceny [%]

- *fakt_průběh*
  (f_průběh_prodejní_cena, f_průběh_vyvolávací_cena, f_průběh_kup_teď_cena, f_průběh_příhoz)

Datový sklad tedy obsahuje šest dimenzionálních tabulek a dvě tabulky faktů. Tabulka *fakt_statistika* obsahuje statistiky ohledně jednotlivých aukcí. Tabulka *f_průběh_příhoz* navíc obsahuje informace o průběhu aukcí – jednotlivé příhozy. I když se zdá, že obě tabulky obsahují prakticky totéž až na jeden atribut, jejich rozdělení bylo nutné. Tabulka statistiky v jistém slova smyslu obsahuje „statická“ data, tedy údaje o již ukončených aukcích. Jaká byla vyvolávací cena, za jakou cenu se předmět prodal, v jaké kategorii byl umístěn a tak dále. Naproti tomu tabulka průběh zaznamenává „dynamiku“ jednotlivých aukcí. Tedy, kdo v průběhu konání na aukci přihazoval a kolik. K jedné aukci (jeden záznam v tabulce statistika) tak v této tabulce může existovat záznamů klidně dvacet (dvacet příhozů). Proto sloučení obou těchto informací do jediné tabulky je jak z praktického hlediska, tak z hlediska přehlednosti a jednoduchosti řešení nevhodné.

Další zajímavosti zjištěné při přípravě dat:
Během přípravy dat se již daly odhalit některé zajímavé údaje, například z celkového počtu 493 263 záznamů, má pouze 121 266 (24,6%) unikátní název aukce. Čili 75,4% aukcí má stejný název jako nějaká jiná aukce.

Nejčastěji se vyskytující názvy aukcí (*Tabulka č. 5*):

Název aukce	Počet výskytů
Bankovka	1236
Mince Dánsko	490
Medaile	477
Mince	468
Bankovka Argentina	326

Tabulka č. 5: nejčastěji se vyskytující názvy aukcí

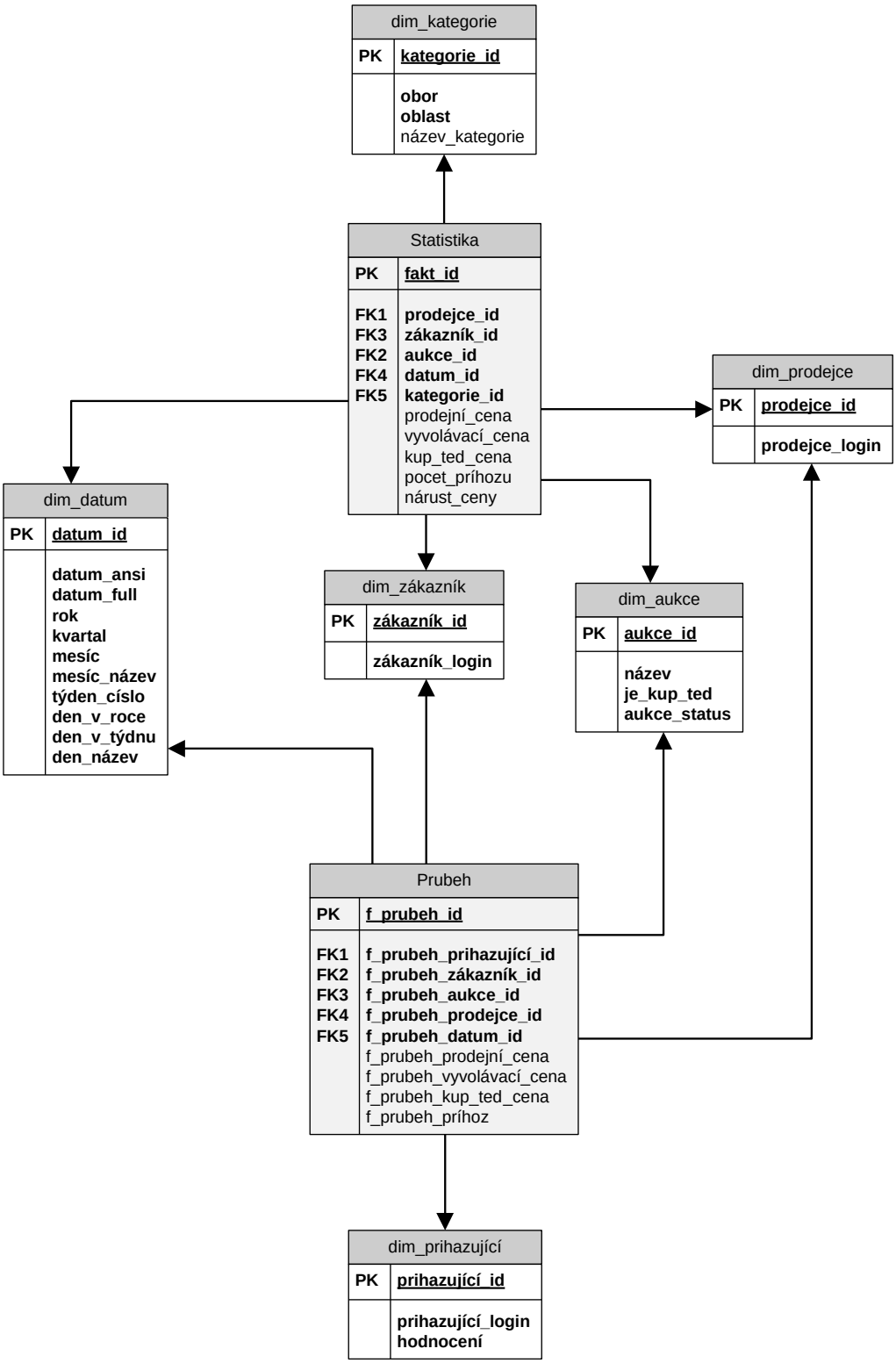
Jelikož v původní databázi jsou atributy pojmenovány anglicky a pro lepší pochopení problematiky byly později uvedeny jejich české ekvivalenty, následuje tabulka (*Tabulka č. 6*), která shrnuje překlad použitých atributů. V následujícím textu již budou používány výhradně české názvy.

Anglický název atributu	Český překlad
Auction_id	Aukce_id
Auction_name	Aukce_Název
Auction_price	Prodejní_cena
Auction_starting_price	Vyvolávací_cena
Auction_is_buy_now	Je_kup_teď
Auction_buy_now_price	Kup_teď_cena
Auction_bid_count	Počet_příhozů
Auction_seller_id	Prodejce_id
Auction_seller_login	Prodejce_login
Auction_bidder_login	Zákazník_login
Auction_bidder_id	Zákazník_id
Auction_status	Aukce_status
User_id	Přihazující_id
User_login	Přihazující_login
User_rating	Hodnocení
Bidding_item_amount	Příhoz
Bid_status	Status_příhozu

Tabulka č. 6: Přehled překladu anglických názvů atributů na české ekvivalenty

7.1.4 Hvězdicové schéma datového skladu

Následující obrázek (*Obrázek č. 14*) zachycuje hvězdicové schéma vznikajícího datového skladu.



Obrázek č. 14: Hvězdicové schéma datového skladu

7.2 Architektura použitého řešení

Tato část textu si klade za cíl přiblížit čtenáři architekturu, která byla použita jak pro vybudování DS, tak následně pro OLAP analýzy až po závěrečné vytvoření reportů. Pro implementaci Business Intelligence byla zvolena platforma Microsoft SQL Server 2005, ta společně s řadou nástrojů od společnosti Microsoft nabízí ucelené a funkční řešení. Jako jednu z výhod lze rovněž brát fakt, že veškeré nástroje pocházejí od jednoho výrobce, tím pádem vzájemná integrace a spolupráce, jak technologií, tak nástrojů by měla být jednodušší.

Implementaci BI na platformě Microsoft SQL Server 2005 lze na nejvyšší hierarchické úrovni rozdělit na tři částečně nezávislé a částečně na sebe navazující bloky [16]:

- **Integrace** – získání dat z různých nehomogenních prostředí a jejích případná transformace
- **Analýza dat** – v této fázi dochází k obohacení dat o výsledky analýz, predikce z data miningu a podobně. Tímto se informace stávají cennými pro podporu rozhodování.
- **Reporty** – slouží pro zpřístupnění dat a výsledků analýz uživatelům ve vhodné formě a obsahu

Situaci lépe vyjadřuje následující schéma (*Obrázek č. 15*):

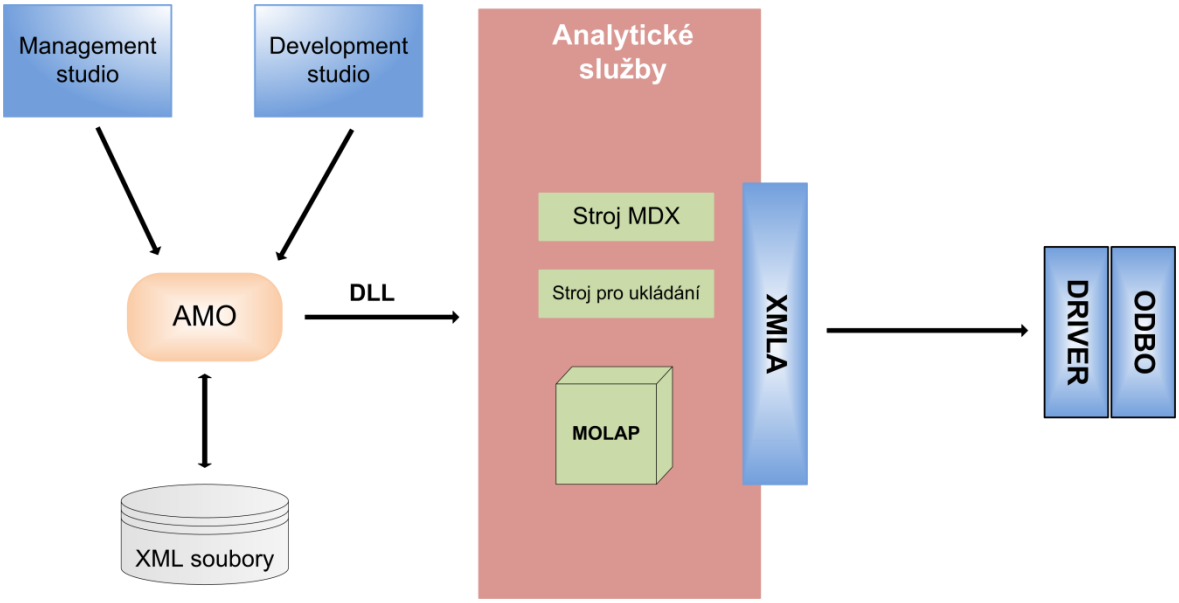
Obrázek č. 15: Základní schéma BI na platformě MS SQL Server 2005

7.2.1 Architektura analytických služeb MS SQL Serveru 2005

V této verzi SQL Serveru se zdrojové soubory obsahující definice objektů OLAP neukládají do speciálního úložiště - tak tomu bylo v předešlé verzi (SQL Server 2000). Nyní se tyto informace uchovávají ve formě XML dokumentů. Přístup k nim je možný pomocí MS Analysis Management Objects (AMO).

Pro správu databází, analytických služeb a analytických objektů slouží aplikace *MS SQL Server Management Studio*. Pro vytváření business intelligence objektů – integrace, analýzy a tak dále se používá aplikace Microsoft *Business Intelligence Development Studio* (BIDS). Jedná se o aplikaci postavenou nad aplikací Microsoft Visual Studio 2005. BIDS má stejné uživatelské rozhraní jako Visual Studio, pouze je omezené na vývoj BI aplikací.

Knihovna analytických objektů AMO posílá požadavky na blok analytických služeb (*Obrázek č. 16)*. Modul MDX má na starosti různé kalkulace. Jako komunikační rozhraní je používáno XMLA (XML for Analysis), jde o jediné systémové rozhraní, pomocí kterého například Excel od verze 2007 komunikuje s analytickými službami. Aplikace, které toto rozhraní nepodporují, mohou s analytickými službami komunikovat pomocí rozhraní ODBO (OLE DB for OLAP). Na straně klienta existuje totiž ovladač, který transformuje volání rozhraní ODBO na volání nativního rozhraní XMLA.



Obrázek č. 16: Schéma architektury analytických služeb

XML for Analysis (XMLA)

XMLA je standard pro komunikaci s analytickými službami. Je spravován mezinárodním orgánem XMLA Council [11]. Tento orgán byl založen společnostmi Microsoft, Hyperion a SAS. Standard XMLA je založen na existujících a osvědčených standardech webových služeb (XML, SOAP). Technologicky nahrazuje starší rozhraní OLEDB for OLAP a OLEDB for data mining. Tento standard je v analytických službách MS SQL 2005 podporován nativně. Je to jediné rozhraní pro přístup k analytickým službám, takže všechny aplikace komunikují s analytickými službami prostřednictvím XMLA (například Excel).

7.3 Integrace a zavádění dat (etapa ETL)

Úkolem této etapy je načíst, transformovat a zavést zdrojová data. Výsledkem pak jsou naplněné datové struktury datového skladu. Výslednou datovou strukturu si lze prohlédnout na *obrázku č. 14*, který představuje hvězdicové schéma DS. Kdyby měla být tato činnost zařazena v rámci základní struktury použité architektury (*Obrázek č. 15*), jedná se o první etapu – Integraci a úkoly s ní spojené. Pro tyto potřeby MS SQL Server 2005 nabízí tzv. *integrační služby*, pomocí kterých dochází k transformaci a posléze zavedení dat.

Nástroj BIDS v rámci integračních služeb nabízí celou řadu sofistikovaných prvků tzv. *Tasks (úlohy),* jde o samostatně vykonatelné, jednoduché částečné úlohy podílející se na komplexním procesu transformace. Tasks lze přirovnat k příkazům libovolného programovacího jazyka. Tyto úlohy lze vzájemně propojovat, vytvářet cykly, podmínkové větve a ve výsledku tak vytvořit velice sofistikovaný projekt pro zpracování dat z heterogenních zdrojů.

Integrační služby obsahují nespočet komponent (tasks), jak na procesní úrovni, tak na úrovní datové (transformace a zavedení dat). Záleží pouze na konkrétních potřebách a požadovaném výsledku. Namátkou například komponenty umožňující zkopírovat celý sloupec, odvodit sloupec nový na základě hodnoty určitého atributu, rozdělit datový tok na základě vstupní podmínky a celá řada dalších. V rámci jednoho projektu je možné mít několik balíčků (*packages*), kdy jeden balíček představuje určitou činnost – vzájemně propojené úlohy. Jelikož cílem této práce není popisovat architekturu MS SQL 2005, nebudou jednotlivé komponenty blíže specifikovány. Následující text pouze ukazuje použití některých těchto možností ve vztahu k vznikajícímu datovému skladu. Zájemcům o popis možností integračních služeb lze doporučit publikaci od L. Lacka [15], která se detailně věnuje problematice Business Intelligence v MS SQL Serveru 2005. Cílem této kapitoly je ukázat a popsat jakým způsobem byl prostřednictvím integračních služeb naplněn datový sklad.

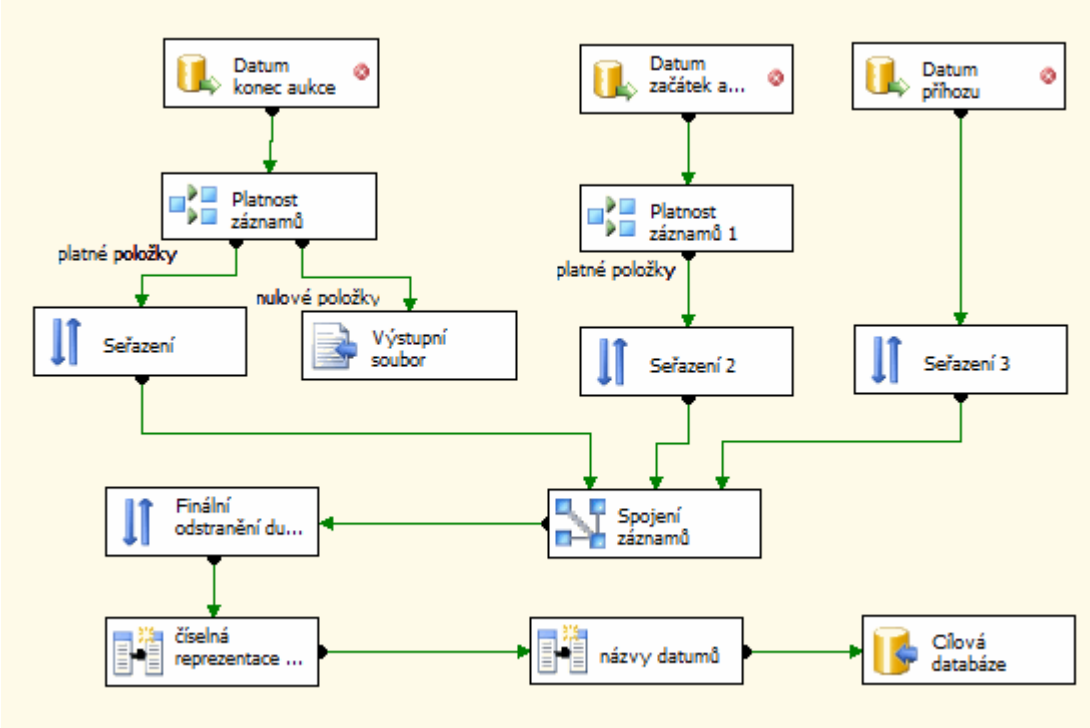
7.3.1 Integrační projekt DS

Projekt integračních služeb *Auction Integration*, který lze nalézt na přiloženém CD obsahuje veškeré úkony potřebné pro naplnění požadovaných struktur. Jeho součástí je několik balíčků:

- *auction_dimension* – naplnění dimenzionální tabulky dim_aukce
- *category_dimension* – naplnění dimenzionální tabulky dim_kategorie
- *cr_sr_br_dimensions* – naplnění dimenzionálních tabulek dim_zakaznik, dim_prodejce, dim_prihazujici
- *date_dimension* – naplnění dimenzionální tabulky dim_datum
- *fact_stats_push_data* – naplnění tabulky faktů statistika
- *fact_duration_push_data* – naplnění tabulky faktů prubeh
- *tempDataA, tempDataB* – přenos zdrojových dat ze zdrojového prostředí do prostředí MS SQL

Poznámka:
Zdrojová data byla uložena v MySQL databázi. Pro naplnění faktů je však potřeba pomocí SQL dotazu spojit dimenzionální tabulky se zdrojovými daty. Jelikož však dimenzionální tabulky již byly uloženy v MS SQL databázi, nabízelo se jako nejjednodušší řešení přenést zdrojová data také do prostředí MS SQL a poté realizovat spojení tabulek v rámci jedné databáze na stejné platformě.

Pomocí těchto balíčků jsou tedy naplněny datové struktury datového skladu. Následující schéma (*Obrázek č. 17*) znázorňuje diagram pro naplnění datumové dimenze. Lze na něm vidět, jaké komponenty byly použity, jakým způsobem jsou vzájemně propojeny a rovněž, jaké operace a transformace se s daty prováděly.



Obrázek č. 17: Ukázka schématu pro naplnění datumové dimenze (Integrační služby MS SQL 2005)

Na obrázku výše je vidět několik komponent integračních služeb, které byly použity, rovněž je znázorněn celý proces od prvotního načtení dat (vrchní 3 komponenty) přes nejrůznější transformace až po finální naplnění dimenzionální tabulky.

Obdélníky označené *Datum konec aukce*, *Datum začátek aukce* a *Datum příhozu* představují jednolité atributy ze zdrojové databáze, které jsou pomocí komponenty *DataReader Source* načteny a dále upravovány. Business Inteligence Development Studio ve výchozí instalaci bohužel nepodporuje připojení k MySQL databázím. Je nutné doinstalovat **Connector/.NET 6.0,** což je nástroj přímo od vývojového týmy MySQL a umožňuje přistupovat k MySQL databázím prostřednictvím ADO.NET přímo z projektů integračních služeb.

V následujícím kroku dochází k odfiltrování neplatných položek (záznamů s hodnotou NULL). Volitelně lze tento výstup přesměrovat například do souboru či databáze pro případ pozdější manipulace s nimi. V dalším kroku dochází k seřazení položek – od nejstaršího data po nejnovější, poté jsou všechny hodnoty těchto tří atributů spojeny do jednoho. Poté je nutné ze záznamů odstranit duplicity, protože jednotlivé dimenze nesmí obsahovat redundantní data. K tomuto kroku dochází v části označené *finální odstranění duplicit*.

V dalším bloku označeném *číselná reprezentace data* jsou získány jednotlivé datumové údaje (den, měsíc, kvartál, rok atd.) a to pomocí vestavěných datumových funkcí, které integrační služby nabízí pro práci s datumovými údaji (*Obrázek č. 18*).

Derived Column Name	Derived Column	Expression
den	<add as new column>	DAY(auction_ending_time_converted)
den_v_tydnu	<add as new column>	DATEPART("weekday",auction_ending_time_converted)
den_v_roce	<add as new column>	DATEPART("dayofyear",auction_ending_time_converted)
mesic	<add as new column>	MONTH(auction_ending_time_converted)
tyden_cislo	<add as new column>	DATEPART("week",auction_ending_time_converted)
kvartal	<add as new column>	DATEPART("quarter",auction_ending_time_converted)
rok	<add as new column>	YEAR(auction_ending_time_converted)

Obrázek č. 18: Vytvoření atributů datumové dimenze

Předposledním krokem v procesu naplnění datumové dimenze je krok, ve kterém jsou vytvořeny názvy jednotlivých datumových položek – názvy dnů v týdnu a měsíců. Jak tato činnost, tak činnost předešlá – extrahování zdrojového data je prováděna pomocí komponenty *derived column*, která umožňuje vytvořit nový sloupec odvozený z dostupných hodnot a to za pomocí nejrůznějších funkcí. V případě vytváření názvů je využíváno hodnot z předešlého kroku a jednoduché logické podmínky (*Výraz č. 1*):

```
Den_nazev = den_v_tydnu == 1 ? "Pondělí" :
(den_v_tydnu == 2 ? "Úterý" : (den_v_tydnu == 3 ? "Středa" :
(den_v_tydnu == 4 ? "Čtvrtek" : (den_v_tydnu == 5 ? "Pátek" :
(den_v_tydnu == 6 ? "Sobota" : "Neděle")))))
```

Výraz č. 1: Logická podmínka pro vytvoření názvů dnů

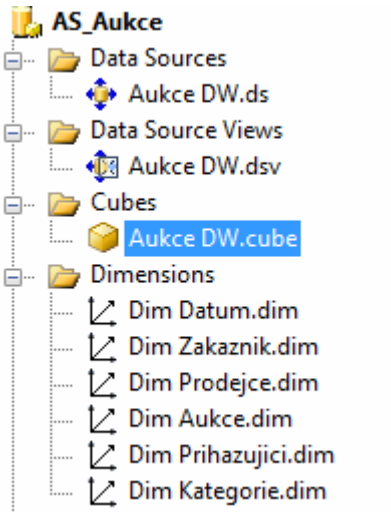
Poslední krokem v celém procesu je uložení všech těchto hodnot do cílové databáze – do tabulky *dim_datum*. Tato činnost je vykonána pomocí komponenty *OLE DB Destination*. V rámci nastavení této komponenty je nutné nastavit mapování zdrojových atributů na cílové.

Stejným postupem jsou naplněny všechny dimenzionální tabulky i tabulky faktů. V některých případech je celý proces značně jednoduší – není potřeba tolik transformací a úprav. Datumová dimenze společně s tabulkami faktů patří v celém integračním procesu k těm nejkomplexnějším, co se týče použitých transformací.

7.4 Vytvoření OLAP kostky

Po etapě ETL, kdy již jsou data zavedená do databáze datového skladu, přichází na řadu další část v procesu budování DS. Nyní je nutné vytvořit OLAP kostku, pomocí které budou následně data analyzována. Pro vytvoření kostky a následné analýzy slouží na platformě MS SQL Server 2005 tzv. *analytické služby.* Jde o typ projektu (podobně jako v případě integračních služeb v předchozí kapitole), pomocí kterého se v BIDS kostky vytváří a následně je možné provádět také požadované analýzy.

Vytvoření kostky se provádí v několika na sebe navazujících krocích. Nejprve je nutné specifikovat datový zdroj (*data sources*), následně pohled na zdrojová data (*data source views*), teprve poté je možné přistoupit k vytvoření samotné kostky (*cubes*). Během této činnosti dochází k vytvoření jednotlivých dimenzí (*dimensions*), se kterými lze následně dále pracovat. Posloupnost jednotlivých kroků přehledně zachycuje *solution explorer* v BIDS, kde lze ihned identifikovat, v jakém pořadí mají být jednotlivé operace prováděny (*Obrázek č. 19*).



Obrázek č. 19: Ukázka struktury projektu analytických služeb

Pro každou tuto aktivitu nabízí BIDS přehledného průvodce, pomocí kterého jsou jednotlivé části vytvořeny. V případě vytváření kostky nabízí průvodce nejvíce možnosti nastavení. Snaží se automaticky rozeznat, které tabulky patří mezi fakta a které mezi dimenze. Automaticky nabídne možnost vytvoření hierarchií tam, kde to má smysl. Samozřejmě je možné dodefinovat ručně další vhodné hierarchie a se vším manipulovat i později po vytvoření kostky. V tomto případě byly vytvořeny dvě hierarchie v datumové dimenzi (*rok-kvartál-měsíc* a *rok-číslo týdne-den*), jedna hierarchie byla také vytvořena v dimenzi kategorie (*obor-oblast-název kategorie*).

Vytvoření samotné kostky je poté závislé na počtu dat a výkonu počítače. Po vytvoření kostky je možné s ní dále manipulovat – vytvářet nová měřítka (agregační funkce aplikovaná na fakta), vytvářet perspektivy, prohlížet si hvězdicové schéma datového skladu a tak dále. Pomocí vestavěného prohlížeče lze přímo v prostředí BIDS prohlížet samotná data agregována podle určitých dimenzí a udělat si tak obrázek o možných výsledcích a hlavně vyzkoušet zda datový sklad, tak jak je navržený, poskytuje nějaké relevantní výsledky nebo je potřeba provést určité úpravy.

7.5 OLAP analýza a reporty

V momentě, kdy již je vytvořená kostka, přichází na řadu samotná OLAP analýza. Následně je nutné výsledky analýzy doručit klíčovým uživatelům. Toho se docílí pomocí tzv. reportů. Architektura MS SQL Server 2005 nabízí tvůrcům datového skladu široký výběr možností, co se týče aplikací, které lze použít pro reporty (česky lze report nazvat například českým ekvivalentem zpráva, v následujícím textu je však používán termín report). MS SQL Server 2005 dokonce nabízí vlastní reportovací server, který je možné použít.

Pro potřeby této práce byl jako reportovací nástroj použit Microsoft Excel ze sady Microsoft Office 2007. Ten, jak je popsáno v části věnující se použité architektuře, komunikuje s analytickými službami prostřednictvím nativního rozhraní XMLA. Použití Excelu jako reportovacího nástroje sebou přináší celou řadu výhod při nasazení v obchodním prostředí. Jde o aplikaci, kterou klíčoví uživatelé s velkou pravděpodobností znají a umějí používat. Odpadá tak nutnost školit uživatele kvůli nové aplikaci. Čímž se je možné snížit náklady, navíc není nutné investovat do pořízení nového softwaru. Tento kancelářský balík je de facto standardem v oblasti kancelářských aplikací, proto je dostupný prakticky v každé společnosti. Další výhodou je, že nehrozí jakýsi „psychický blok“ ze strany uživatele z důvodu, že nová aplikace je složitá na ovládání a má strach, aby se s ní naučil pracovat. V neposlední řadě použití nástrojů výhradně od jedné společnosti může vést k větší spolehlivosti, jelikož zde existuje větší pravděpodobnost, že by vše mělo fungovat správně.

Připojení k datovému skladu z Excelu je velice pohodlné. Slouží k tomu opět průvodce, kterého lze vyvolat na kartě *Data* a zde je položka *Načíst externí data z jiných zdrojů*. V případě datového skladu je nutné zvolit položku: *ze služby pro analýzu*. Průvodce je velice jednoduše navržen, stačí pouze zvolit název SQL Serveru a název databáze, která obsahuje požadované údaje. Dále již průvodce nabídne samotnou OLAP kostku, případně perspektivy, které je možné vybrat pro načtení. Data lze zobrazit buď pomocí kontingenčního grafu, nebo pomocí kontingenční tabulky.

Excel se v takovém případě připojí k OLAP serveru jako klient tzn., že pro práci s daty je potřeba být neustále připojen. V takovém případě je možné přímo v Excelu s daty on-line manipulovat a vytvářet libovolné exporty (*Obrázek č. 20*). Není poté problém dosažené výsledky ihned integrovat do jiných firemních dokumentů, prezentací, výroční zpráv a podobně a to v libovolné formě – tabulka, graf. Excel dokonce umožňuje práci s OLAP kostkou v režimu off-line. V takovém případě se OLAP kostka uloží jako soubor s příponou .CUB na lokální počítač a pak je možné s daty pracovat i po odpojení od OLAP serveru. Tento soubor obsahuje všechny potřebné údaje (fakta, dimenze). pro vytváření kontingenčních tabulek a grafů. V případě rozsáhlých databází a velkého objemu dat je však nutné počítat s vysokými nároky souboru na úložný prostor. Každá dimenze exponenciálně zvyšuje nároky na diskovou kapacitu. Proto v případě, že je nutné s daty pracovat offline je vhodné u velkých databází dostatečně rozvrhnout, které dimenze jsou pro danou situaci potřeba a uložit pouze je.

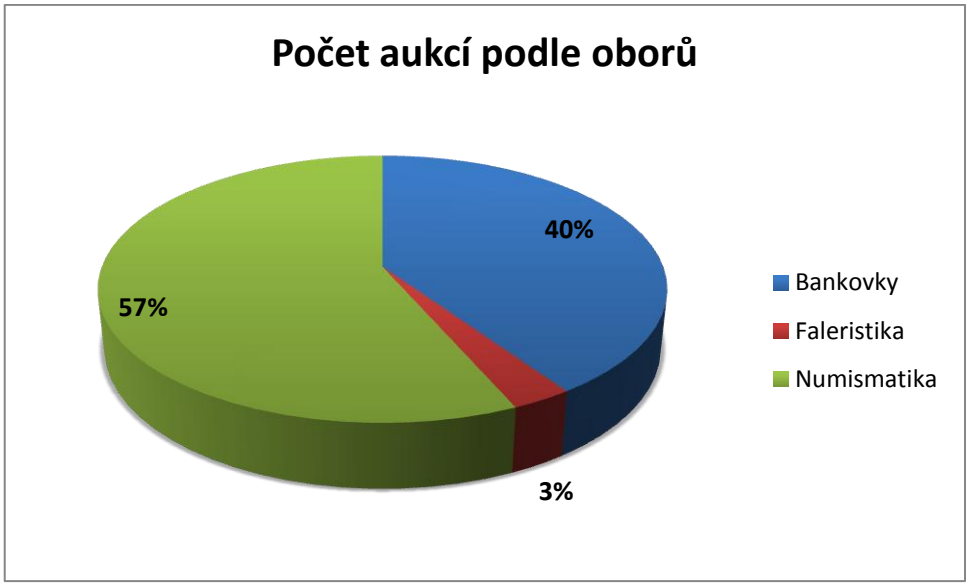
Počet aukcí	Popisky sloupců								
	⊟2008						Celkem z 2008	⊞2009	Celkový součet
	⊞2	⊞3	⊟4			Celkem z 4			
Popisky řádků			Listopad	Prosinec	Říjen				
⊞Bankovky		4986	5052	5350	4393	14795	19781	169021	188802
⊞Faleristika				620		620	620	13299	13919
⊟Numismatika	97	9378	8263	7696	6914	22873	32348	231667	264015
⊞Afrika								7209	7209
⊞Antické								1857	1857
⊞Asie								11288	11288
⊞Austrálie a Oceánie								2606	2606
⊟Česko	60	6398	5571	5038	4554	15163	21621	93360	114981
1918-1945								21193	21193
1945-1992 oběžné								22714	22714
1945-1992 pamětní		2185	2116	1686	1743	5545	7730	20958	28688
Od 1993 oběžné		1065	1379	1569	1155	4103	5168	9581	14749
Od 1993 pamětní	60	3148	2076	1783	1656	5515	8723	18914	27637
⊞Euro mince								8578	8578
⊞Evropa								45314	45314
⊞Medaile								4718	4718
⊞Německo								13820	13820
⊞Nouzové mince								96	96
⊞Rakousko-Uhersko								7353	7353
⊞Severní Amerika								2134	2134
⊞Slovensko		878	767	870	692	2329	3207	5123	8330
⊞Středověk								6865	6865
⊞Tolary								1607	1607
⊞USA								6295	6295
⊞Zlaté mince	37	2102	1925	1788	1668	5381	7520	13444	20964
Celkový součet	97	14364	13315	13666	11307	38288	52749	413987	466736

Obrázek č. 20: Ukázka exportu z aplikace Microsoft Excel

Na obrázku č. 20 je zachycen export z aplikace Microsoft Excel. Konkrétně zachycuje počet aukcí v jednotlivých kategoriích za časové období. S kontingenční tabulkou lze libovolně manipulovat – rozbalovat a zabalovat jednotlivé sloupce až na elementární úroveň, přidávat další popisné atributy (dimenze). V tomto případě například lze přidat ještě jednotlivé prodejce a sledovat tak počet aukcí v jednotlivých kategoriích za časové období a pro jednotlivé prodejce. Vše se samozřejmě děje on-line a data se načítají přímo z OLAP serveru. Na obrázku výše jsou zachyceny dvě hierarchie. Jedna časová: *rok – kvartál – měsíc* a hierarchie kategorií: *obor – oblast – název kategorie.*

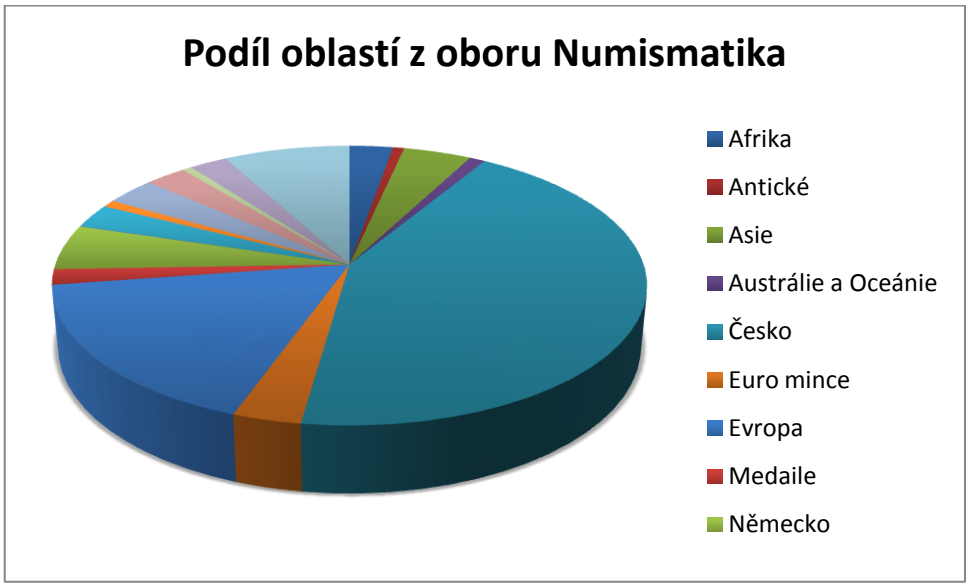
7.6 Zhodnocení výsledků OLAP analýzy

V rámci této práce byl samozřejmě podle postupů, které tato práce zmiňuje vybudován datový sklad. Cílem této kapitoly je zhodnotit a vhodně okomentovat výsledky OLAP analýzy a pokusit se nastínit využití nově nabytých informací v obchodním prostředí. Výsledky jsou prezentovány pomocí přehledných grafů a komentářů.



Graf č. 1: Celkový počet aukcí, podle jednotlivých oborů

Na *grafu č. 1* je znázorněn celkový počet aukcí ve vztahu k jednotlivým oborům. Naprosto zanedbatelně v porovnání s ostatními působí obor *Faleristika*. V oblast Numismatika patří jednoznačně mezi nejoblíbenější. Celkem je v aukčním systému evidováno 264 015 aukcí v této oblasti. Naproti tomu *graf č. 2* zachycuje jednotlivé oblasti náležící oboru Numismatika, tedy oboru s největším počtem aukcí. Zde je patrná naprostá dominace aukcí náležících oblasti *Česko* – celkem 114981 aukcí, čili zboží této kategorie je jednoznačně nejpopulárnější v celém aukčním systému. Dalšími oblastmi s výraznějším podílem jsou oblasti *Evropa* a *Zlaté Mince*. Aukce náležící oblasti *Česko* mají celkem 24, 6 % podíl ve vztahu k celému aukčnímu systému. Celkem je v systému evidováno **466 736** aukcí. Naopak na opačném konci, co se týče počtu aukcí v oboru Numismatika, stojí oblast *Nouzové mince* s celkovým počtem 96 aukcí. V případě oblasti Česko lze jít ještě dále dopodrobna. V takovém případě je nejoblíbenější kategorie z oblasti Česko kategorie *pamětní mince z období 1945-1992* s celkovým počtem 28 688 aukcí.



Graf č. 2: Podíl oblastí z oboru Numismatika

Počet aukcí z časového hlediska

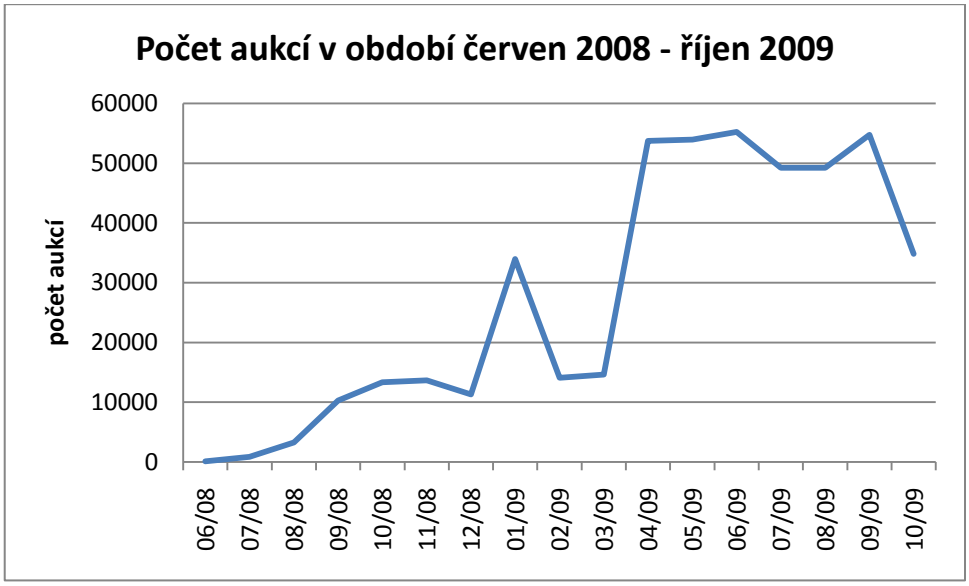
Následující *graf č. 3* zachycuje vývoj počtu aukcí za specifikované časové období. Z grafu je patrný strmý nárůst od dubna 2009, kdy počátkem léta dochází k mírnému poklesu a celkovému ustálení, na podzim naopak dochází opět k rapidnějšímu poklesu. Lze z toho vydedukovat například, že během letních měsíců je mezi lidmi větší zájem o danou komoditu. Naproti tomu však je pravděpodobné, že v letních měsících budou lidé trávit u počítačů, a tedy i Internetu méně času než v jiných ročních obdobích. Zda je tato hypotéza pravdivá či nikoliv může pomoci odhalit data mining, proto je tento problém zmiňován také později v části věnované dolování znalostí z dat.

Další zajímavou informací spojenou s časem je den, kdy končí aukce. Nejvíc aukcí končí v neděli (*103 531*) následováno sobotou (*69 827*). Z toho lze vydedukovat, že prodejci již mají vysledováno, že přes víkend lidé mají přeci jenom více času, tak své aukce zveřejňují tak, aby většinou končily přes víkend. Jelikož opravdový „boj" o jednotlivé položky se odehrává vždy až v posledních minutách či sekundách před skončením aukce.

Další informace o počtu aukcí z časového hlediska:

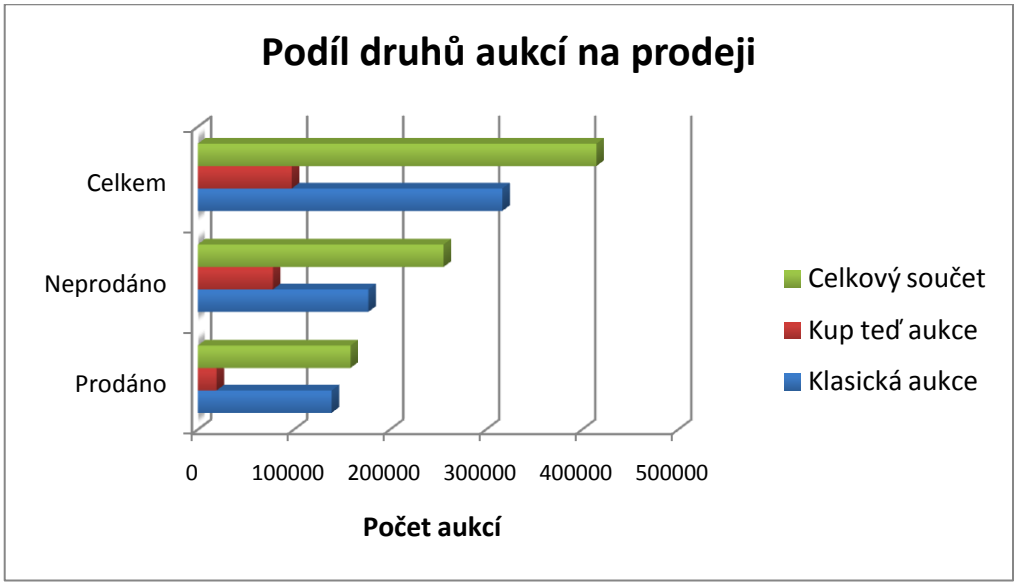
- V druhé polovině roku 2008 bylo vytvořeno celkem 52 749 aukcí.
- Přibližně stejný počet aukcí (53 746) bylo v roce 2009 vytvořeno za měsíc (červen)
- Nejvíc aukcí vzniklo v květnu 2009 a to 55 247
- Za období leden – říjen 2009 vzniklo celkem 413 987 aukcí
- Nejúspěšnější období je 2. kvartál roku 2009 (celkem 162 939 aukcí)
- Nejúspěšnější den z hlediska počtu aukcí je neděle 11. Října 2009 (3 848 aukcí)

Zajímavou informací je zejména údaj, že za deset měsíců roku 2009 vzniklo pomalu 8x víc aukcí než za druhou polovinu roku 2008. Z toho lze usuzovat narůstající oblíbenost tohoto oboru, případně ochota lidí nakupovat toto zboží prostřednictvím aukčního portálu.



Graf č. 3: Vývoj počtu aukcí za období červen 2008 – říjen 2009

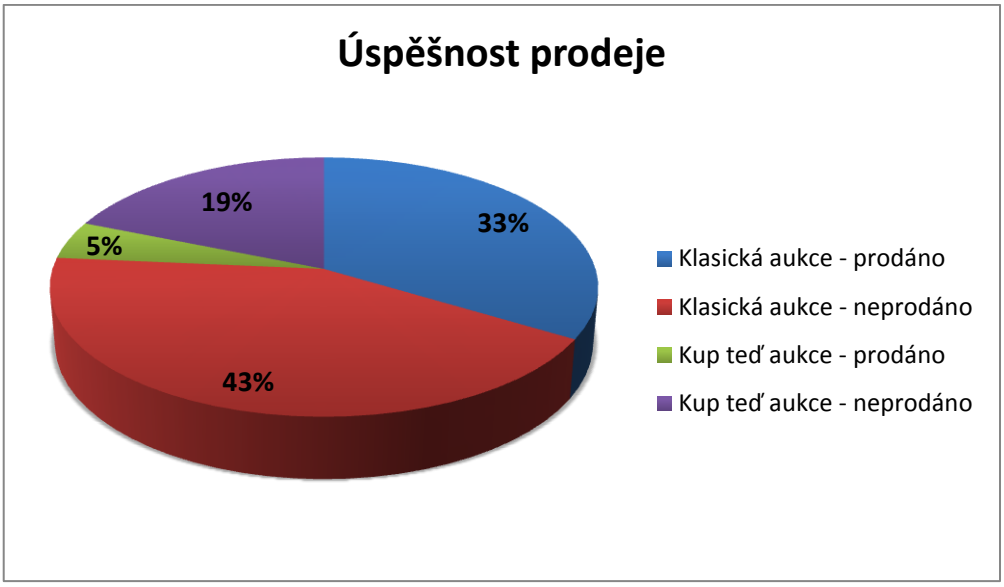
Na grafu č. 4 je zachyceno rozložení klasických aukcí s příhozy a aukcí typu kup teď. Je patrné, že prakticky tři čtvrtiny z celkového počtu, přesněji 75,5 % tvoří klasické aukce, kde uživatele přihazují své částky a navyšují tak celkovou cenu zboží. Aukce s pevnou cenou (kup teď) tedy nejsou mezi sběrateli mincí a bankovek příliš oblíbené. Kromě této informace graf znázorňuje také, jak úspěšné aukce jsou. Opět je zcela patrné, že větší polovina zboží se neprodá. Z celkového počtu 466 736 aukcí neskončí prodejem zboží 54,8 % aukcí. Z toho 16,8 % tvoří aukce kup teď, zbytek jsou klasické aukce. Naopak úspěšným prodejem skončí celkem 34,1 % aukcí. To lze označit za velice příznivé, jelikož prakticky každá třetí aukce skončí prodejem. U bezmála půl milionu případů už jde o zajímavé číslo. Na těchto úspěšně ukončených aukcích se aukce typu kup teď podílejí 12,4 %. Nejvíce úspěšných i neúspěšných aukcí je logicky v oboru Numismatika, která obsahuje nejvíc aukcí. Konkrétně v této kategorii skončí úspěšným prodejem celkem 36,2 % aukcí.



Graf č. 4: Podíl jednotlivých druhů aukcí na prodeji

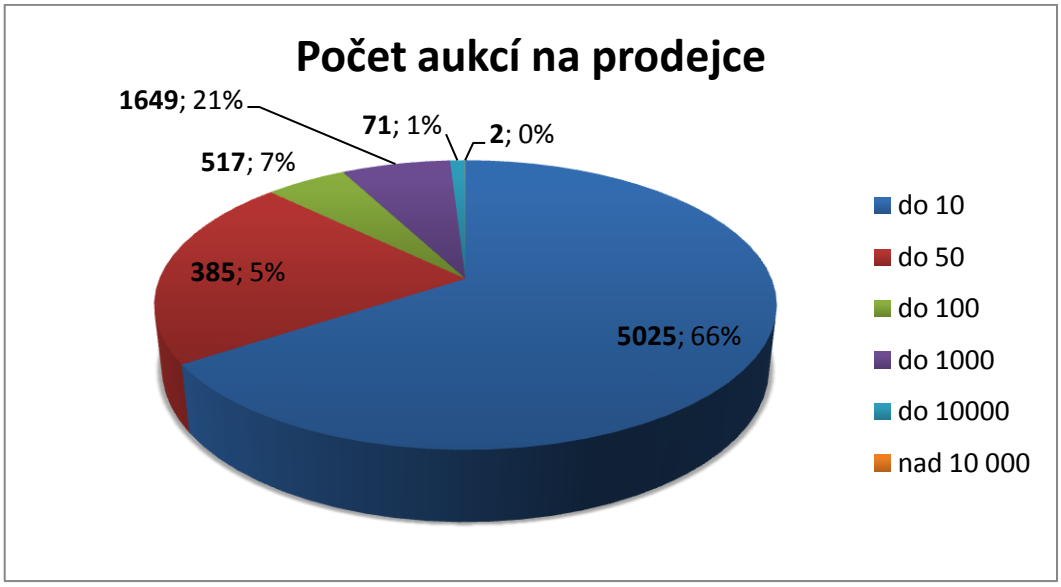
Lépe tuto situaci určitě zachycuje graf č. 5, který ukazuje úspěšnost prodeje vzhledem k typu akce. Na grafu si lze všimnout mírně odlišných hodnot, než které jsou uvedeny v minulém odstavci. To je způsobeno tím, že v minulém odstavci jsou jednotlivé počty aukcí porovnávány s celkovým počtem aukcí v systému (466 736). Toto číslo kromě ukončených aukcí, ať prodaných či neprodaných zahrnuje také aukce, které nebyly ukončeny, byly v průběhu zrušeny nebo neměly tuto informaci uvedenu. Jedná se cca o 10 % z celkového počtu záznamů. Naproti tomu graf č. 5 již je vztahován pouze k aukcím, které byly ukončeny. Proto ty mírné odchylky v řádu procent. V situaci zachycené na grafu č. 5 dokonce úspěšnost prodeje stoupá pomalu k 40 %, přesněji celkem 38 % aukcí skončí prodejem. Opět je patrný minimální podíl kup teď aukcí.

Při pohledu na počet aukcí a uživatele, nejaktivnější uživatel má v systému celkem 36 754. Z tohoto počtu však pouze 11,7 % skončilo úspěšným prodejem. 36 620 aukcí tohoto uživatele pochází z oblasti Bankovky. Čili se zde projevuje specializace na určitý segment trhu, který je patrné také u prodejců s menšími objemy. Částečně se toto ukazuje také u některých zákazníků – nakupují více předmětů od jednoho prodejce. Opět se jedná o zajímavou hypotézu, která bude dále prozkoumána v části věnující se dolování znalostí z dat.



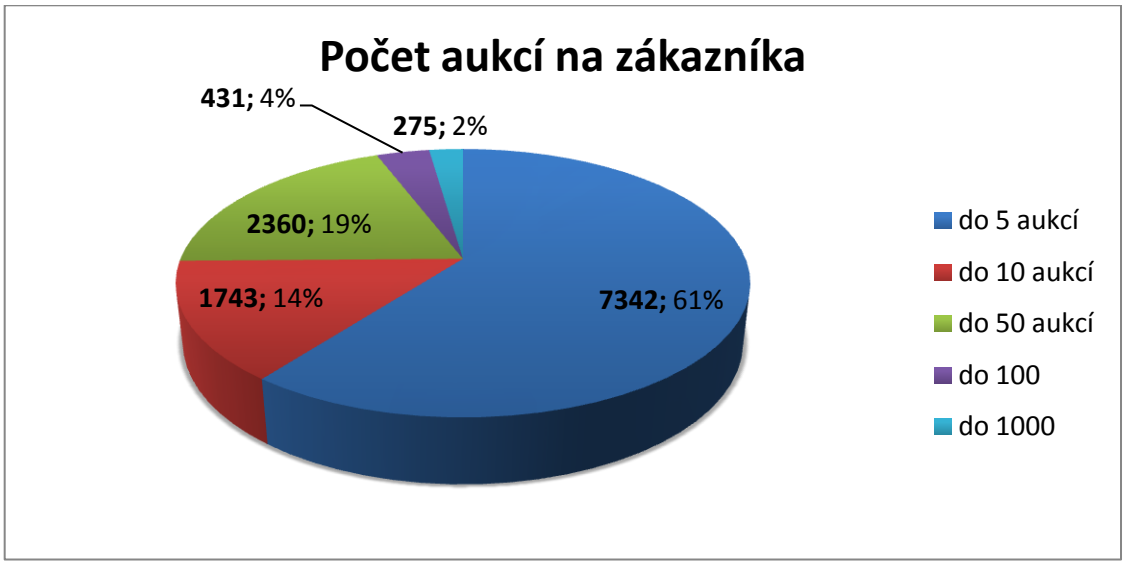
Graf č. 5: Úspěšnost prodeje podle typu aukce

Další dva grafy se věnují počtu aukcí vzhledem k prodejcům respektive zákazníkům. Zachycují tak, jaká je situace na trhu z hlediska počtu aukcí připadajících na prodejce, zákazníka. V případě prodejců (graf č. 6), je patrné, že dvě třetiny z nich vystavili v systému maximálně 10 aukcí. V systému se objevují pouze dva prodejci, kteří za sledované období v systému vystavili více jak 10 000 předmětů. Nezanedbatelnou skupinu dozajista také tvoří prodejci s počtem vystavených aukcí do 10 000 společně se skupinou prodejců vystavujících do 1 000 předmětů. V těchto objemech se dá předpokládat, že se nejedná pouze o hobby daných lidí, ale tito lidé využívají aukční systém jako další obchodní kanál pro distribuci svého zboží. V celkovém objemu se nejedná o nijak zanedbatelný počet (1 722 prodejců, tj. 22 %). Celkem je v systému evidováno 7649 prodejců. Na druhou stranu je také jasně vidět, že pro většinu prodejců (78 %) se s největší pravděpodobností jedná pouze o jejich koníček a aukčního portálu využívají ve formě jakési burzy, kde mohou udat předměty, které třeba již ve své sbírce nechtějí a tak dále.



Graf č. 6: Počet aukcí na prodejce

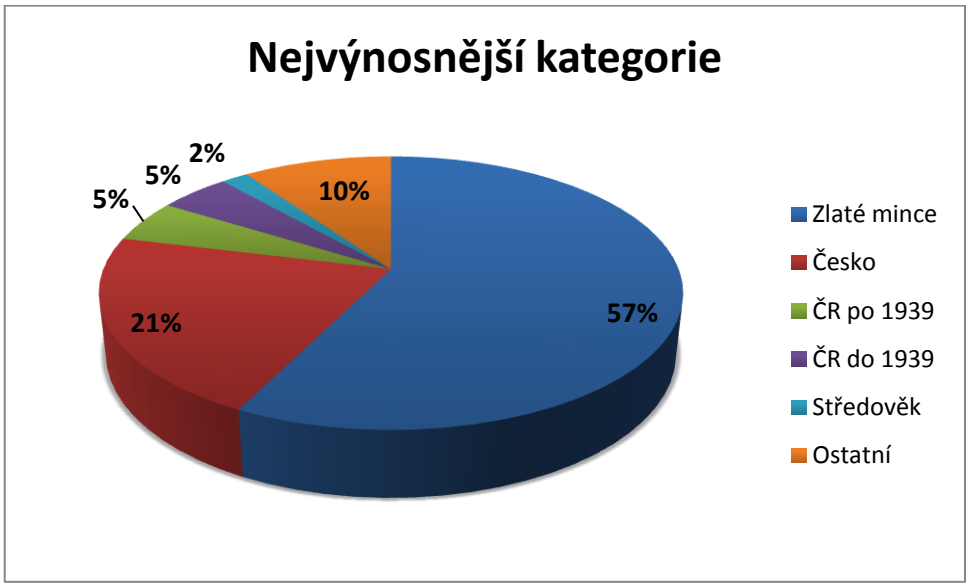
K podobným závěrům lze dospět také v případě zákazníků (graf č. 7). I zde je patrné, že valná většina zákazníků nakupuje pouze menší počet předmětů. 7 342 zákazníků z celkového počtu 12 151 nenakoupila za sledované období více jak 5 předmětů. Pouze 2 % zákazníků nakupují předměty ve velkém – po stovkách kusů.



Graf č. 7: Počet aukcí na zákazníka

Analýza objemu peněžních prostředků v rámci systému

Celkem aukčním systém za sledované období (červen 2008 – říjen 2009) proteklo **144 540 409 Kč.** Jinými slovy bylo prodáno zboží v celkové hodnotě cca 144,5 milionů korun. V rámci celého systému byly zjištěny kategorie, které se na této sumě podílejí největší měrou. Jejich vztah k celkovému objemu peněz znázorňuje graf č. 8. Zcela největší měrou na celkovém objemu peněz se podílí kategorie Zlaté Mince, plných 57% z celkového objemu pocházejí z této kategorie. Zajímavé je také to, že v této oblasti není umístěno největší procento aukcí, tzn. že v této kategorii se prodává zboží za výrazně vyšší ceny než v jiných kategoriích

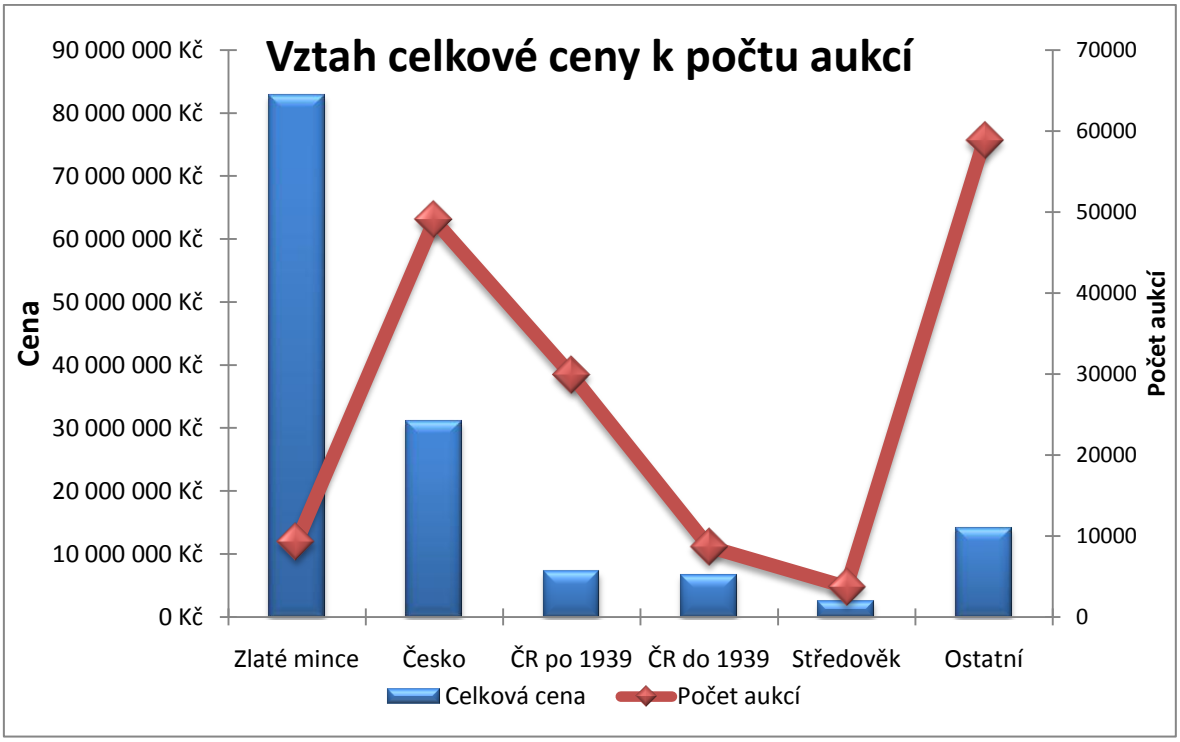


Graf č. 8: Nejvýnosnější kategorie aukčního portálu

Průměrná cena na zboží v kategorii Zlaté Mince činí 8 942 Kč, naproti tomu průměrná cena zboží v rámci celého aukčního systému činí pouze 908 Kč. Tedy je zde patrný výrazný rozdíl. Celkově se také pět nejúspěšnějších kategorií podílí 90 % na celkovém objemu peněz v rámci systému. Navíc v těchto pěti kategorií se rovněž nachází 63 % všech aukcí. Opět se zaběhnutým zvyklostem vyhýbá kategorie Zlaté Mince, která na svůj veliký objem obsahuje pouze 9 273 aukcí, na rozdíl například od kategorie Česko, která obsahuje celkem 48 956 aukcí. Všechny tyto názorně zachycuje tabulka č. 7.

Kategorie	Cena celkem (v Kč)	Počet aukcí	Průměrná cena na aukci (v Kč)
Numismatika	**126 636 724**	**95 579**	**1 325**
Zlaté mince	82 919 977	9 273	8 942
Česko	31 043 310	48 956	634
Středověk	2 554 359	3 702	690
Slovensko	2 530 437	2 845	889
Tolary	1 626 606	678	2 399
Evropa	1 221 680	10 502	116
Euro mince	1 178 181	2 107	559
Rakousko-Uhersko	993 944	2 894	343
Německo	741 216	4 268	174
Medaile	501 831	1 128	445
Antické	498 545	961	519
USA	336 569	2 478	136
Asie	219 834	2 689	82
Afrika	151 749	1 910	79
Austrálie a Oceánie	64 292	670	96
Severní Amerika	49 014	483	101
Nouzové mince	5 179	35	148
Bankovky	**16 722 398**	**59 360**	**282**
ČR po 1939	7 212 332	29 819	242
ČR do 1939	6 709 507	8 659	775
Evropa	1 063 021	6 966	153
Německo	450 122	3 797	119
Amerika	401 044	2 250	178
Rakousko-Uhersko	331 268	977	339
Afrika	219 152	2 003	109
Asie	214 759	4 386	49
Nouzová platidla	91 681	333	275
Austrálie a Oceánie	29 511	170	174
Faleristika	**1 181 288**	**4 168**	**283**
Medaile	1 181 288	4 168	283
CELKEM	**144 540 409**	**159 107**	**908**

Tabulka č. 7: Přehled objemu finančních prostředků v rámci aukčního systému

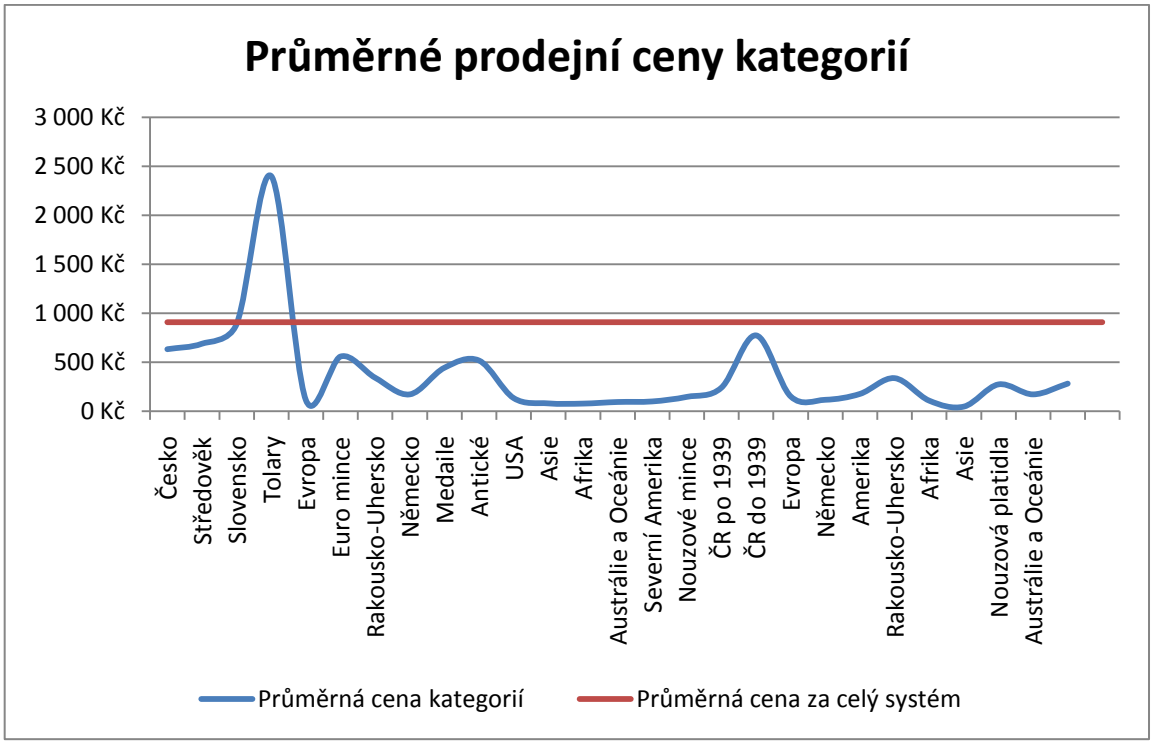


Graf č. 9: Vztah celkové ceny za jednotlivé kategorie k počtu aukcí

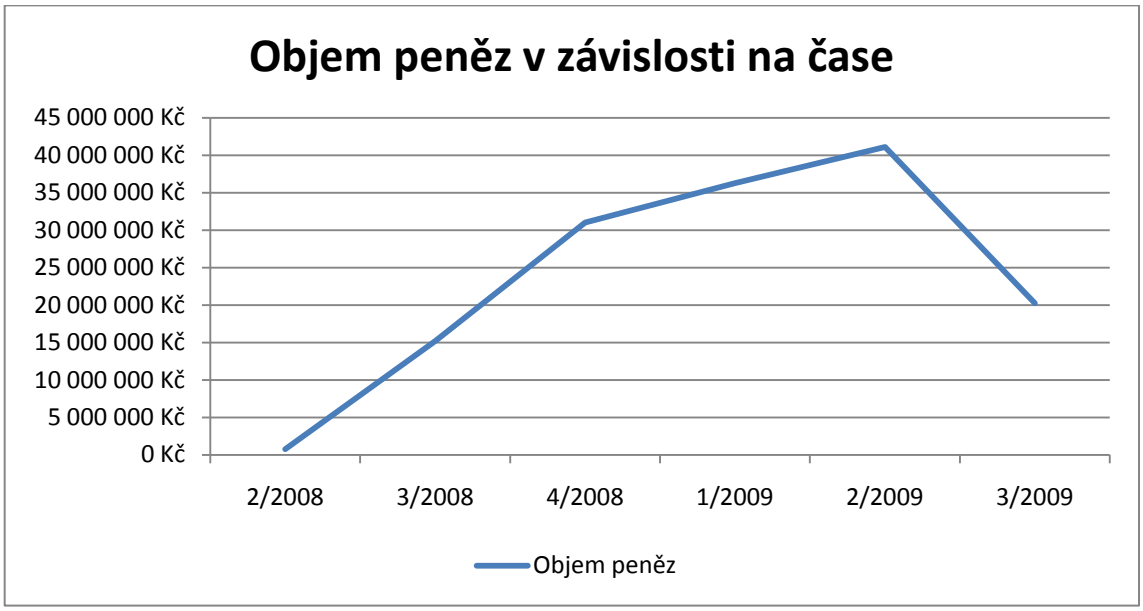
Situaci zachycenou v tabulce č. 7 dále dokreslují následující dva grafy. Graf č. 9 znázorňuje, jaký je vztah počtu aukcí v jednotlivých kategoriích k celkovému objemu peněz, který daná kategorie vygenerovala. Je zde zcela jasně vidět, že kategorie Zlaté mince k svému naprosto unikátnímu objemu v porovnání se zbytkem kategorií, potřebovala velice nízký počet aukcí. Naproti tomu kategorie Česko k získání svého objemu potřebovala cca 5x více aukcí. Nejvíc aukcí bylo samozřejmě umístěno v jiných kategoriích, které jsou sdruženy pod položkou ostatní, ovšem daný peněžní objem této položky není nijak závratný.

Graf č. 10 pak zachycuje vztah průměrné prodejní ceny pro jednotlivé kategorie k průměrné prodejní ceně za celý systém. Z grafu byla záměrně odstraněna kategorie Zlaté mince, jejichž průměrná prodejní cena výrazně převyšuje průměr celého systému a celý graf byl tak výrazně zkreslený. Tuto položku lze považovat za odlehlé pozorování. Z grafu č. 10 je patrné, že průměrná prodejní cena většiny kategorií se pohybuje pod průměrem systému. Čili stanovený průměr vytváří zejména nezobrazená kategorie Zlaté mince se svým průměrem 8 942 Kč. V grafu jsou rovněž zachyceny kategorie s velice nízkou, prakticky nulovou prodejní cenou - *Asie, Afrika, Austrálie a Oceánie*. V současné chvíli se lze pouze domnívat, zda se v těchto kategoriích prodává zboží za takto nízké ceny nebo zda tyto kategorie nepatří mezi příliš oblíbené a mnoho zboží se zde neprodá. Tato hypotéza bude opět dále rozvedena v části věnující se dolování znalostí z dat.

Poslední graf této kapitoly (graf č. 11) zachycuje objem peněz (celková prodejní cena) v závislosti na čase. Opět tento graf potvrzuje vzrůstající oblibu tohoto druhu zboží, zde je to patrné z výše peněz, které systémem protekly za dané období. Navíc je zde patrný také mírný pokles zájmu v letních měsících (3. kvartály jednotlivých let). Možné důvody tohoto poklesu již byly naznačeny.



Graf č. 10: Průměrné prodejní ceny podle kategorií ve vztahu k průměrné ceně celého systému



Graf č. 11: Objem peněz v závislosti na čase

Na závěr pár zajímavostí v bodech:

- Nejvyšší útrata jednoho uživatele celkem činí 2 020 998 Kč
- Nejdražší uskutečněná kup teď aukce byla za 350 000 Kč
- Nejdražší předmět prodaný prostřednictvím klasické aukce stál 710 100 Kč
- Nejvyšší vyvolávací cena prodaného předmětu – 350 000 Kč
- Nejvyšší vyvolávací cena předmětu, který se neprodal – 999 999 Kč

7.6.1 Shrnutí

Během OLAP analýzy bylo zjištěno velké množství zajímavých informací. Rovněž se objevily hypotézy, které mohou být potvrzeny v rámci dolování znalostí z dat. Také je možné, že prezentované výsledky během data miningu budou potvrzeny či více rozvedeny. Z OLAP analýzy jednoznačně vyplynulo, že většina uživatelů aukčního portálu využívá jeho služeb zejména pro potřeby svých zájmů – v tomto případě sběratelství. Ovšem byly také odhaleny skupiny uživatelů, kteří aukční portál využívají jako další trh a nabízí zde své zboží. Dá se s velkou pravděpodobností předpokládat, že se jedná o obchodníky. Zaznamenána byla i rostoucí obliba numismatiky v závislosti na čase, kdy se během roku 2009 prakticky zdvojnásobil objem aukcí v systému. Byly odhaleny kategorie, které patří mezi ty nejoblíbenější mezi uživateli nebo naopak generují největší příjmy. Byla zjištěna celkem vysoká úspěšnost prodeje zboží v rámci této oblasti – 38 % aukcí končí úspěšně – prodejem zboží.

Dosažené výsledky lze ve vztahu k podnikání či trhu jako takovému komentovat ze dvou pohledů:

Z pohledu provozovatele aukčního systému

Provozovatele aukčního portálu může těšit vzrůstající obliba zboží z této oblasti a také relativně vysoké procento úspěšnosti prodeje. Mimo jiné to pro něj znamená nové zákazníky a samozřejmě další provize z prodeje zboží, které jistě už tak jsou dosti vysoké. S prodejci, kteří daný portál využívají ve velké míře (1 000 a více vystavených aukcí), se může dohodnout na bližší spolupráci a nabídnout jim tak zvýhodněné podmínky výměnou za vyšší provizi z prodeje, reklamu v obchodě provozovatele či jiné výhody. U oblíbených kategorií, které obsahují velké množství vystavených aukcí v porovnání s ostatními, může mírně zvýšit poplatky. Uživatelé (prodejce) mírné zvýšení poplatků příliš nerozhodí a i mírné zvýšení poplatků v rámci velkého objemu vystavených aukcí, které v těchto kategoriích jsou, může přinést zajímavé zvýšení příjmů. Další možností je na stránkách se zbožím z těchto kategorií umístit reklamu.

Z pohledu podnikatele, který přemýšlí, zda začít podnikat v této oblasti

V případě podnikatele, který se rozhoduje o tom, zda je výhodné podnikat v oblasti numismatiky sběratelství nebo naopak v případě podnikatele, který v této oblasti již podniká a hledá nové trhy, mohou výsledky této analýzy posloužit jako část analýzy trhu a zcela jistě se z nich dají získat závěry, které mohou pomoci k tomuto strategickému rozhodnutí. Jelikož OLAP analýza odhalila mezi prodejci takové, kteří aukčního portálu využívají jako dalšího trhu pro své zboží, je zde možnost o této variantě uvažovat. Získané výsledky o průměrných cenách jednotlivých druhů zboží podle kategorií, mohou pomoci v rozhodnutí, na který segment se zaměřit.

V souvislosti s tím je také důležitá informace o počtu aukcí v jednotlivých kategoriích a samozřejmě úspěšnost prodeje. Nicméně zcela jistě stále zůstává mnoho informací skryto. Pomocí data miningu mohou být odhaleny další zajímavé skutečnosti, které mohou doposud nabyté informace dále zpřesnit. Dolování znalostí z dat je obsahem následující kapitoly.

8 Data Mining

Data Mining patří v současnosti mezi nejrychleji rostoucí segmenty Business Inteligence a podobně jako OLAP analýzy mají tuto technologii implementovanou ve svých řešeních všichni významní hráči v této oblasti. Data Mining je do češtiny překládán jako metody získávání, dolování znalostí z dat. Jinými slovy jde o další metodiku pro získávání informací pro podporu rozhodování. V některých případech bývá technologie pro dolování dat také označována zkratkou KDD (Knowledge discovery in databases).

Existuje mnoho definic data miningu, Luboslav Lacko ve své knize uvádí tuto [15]:

Data Mining je proces analýzy dat z různých perspektiv a jejich přeměna na užitečné informace. Z matematického a statistického hlediska jde o hledání korelací, tedy vzájemných vztahů nebo vzorů v datech.

Data Mining je tedy z části založen na matematické statistice, heuristických algoritmech, neuronových sítích a jiných pokročilých softwarových technologiích a metodách umělé inteligence. Využití je široké, například v bankovnictví, lékařství, v telekomunikacích apod. Obecně vzato má DM smysl všude tam, kde se předpokládá, že by v datech mohly být ukryty nějaké další skutečnosti a zajímavé informace, které nejsou patrné při klasickém pohledu na data nebo nebyly odhaleny při OLAP analýzách. V současnosti tato oblast BI zažívá boom a velké společnosti napříč průmyslovými odvětvími se snaží implementovat toto řešení do svých procesů.

DM stejně jako datový sklad vyžaduje přípravu dat. Je nutné předzpracování, například vhodně upravit vstupy. Problematice předzpracování dat pro potřeby dolování znalostí z dat se práce zabývá později. Data miningový model nemá smysl aplikovat ani pro potřeby procvičování na databázi, která obsahuje náhodné údaje. Kde se v datech žádné informace nevyskytují, tak je samozřejmě nelze ani vydolovat. V mnoha případech se rovněž může stát, že se žádné nové údaje vydolovat nepodaří, což může mít několik příčin. Buď tam žádné informace nejsou, nebo byla zanedbána etapa přípravy dat či byla vybrána špatná data miningová metoda, případně mohla být nesprávně nastavena. DM je tedy zdlouhavý proces s nejistým výsledkem.

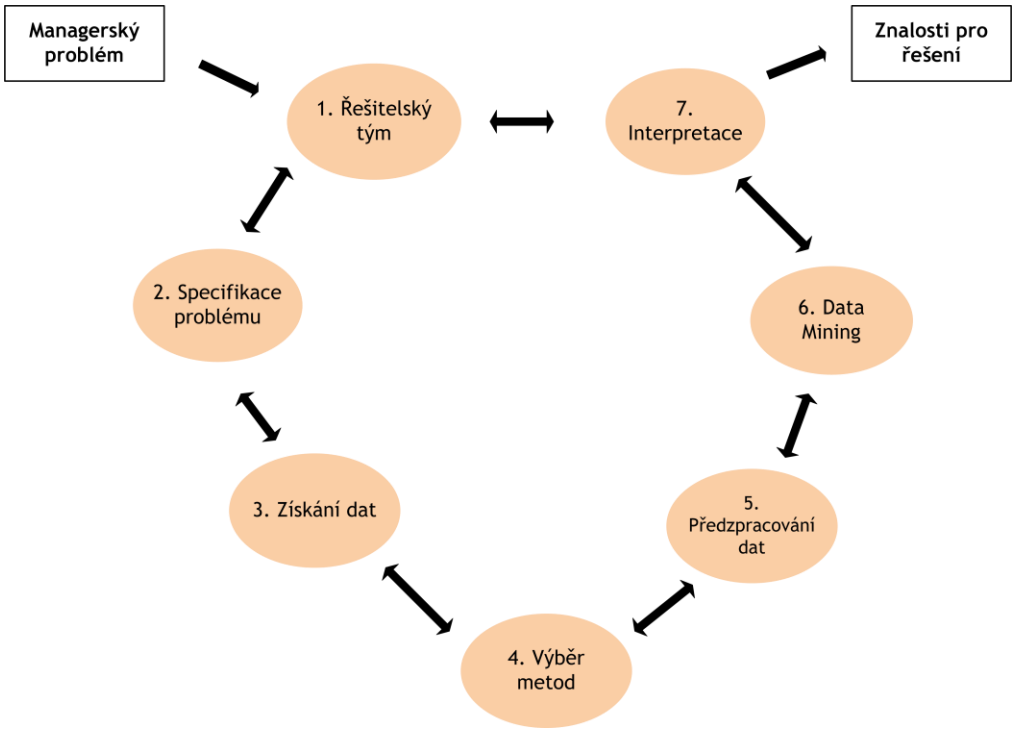
Cílem této kapitoly je uvést čtenáře do problematiky data miningu, vysvětlit základní principy, popsat používané algoritmy.

Rozdíl mezi OLAP analýzou a dolováním znalostí z dat

Může se zdát, že jsou tyto metodiky zaměnitelné, ovšem není tomu tak. Obě se používají ze stejného důvodu - poskytnutí relevantních informací pro podporu rozhodování. Každá však prezentuje jiný typ dat. OLAP pracuje výhradně s agregovanými údaji různých hierarchických stupňů. Výsledky prezentuje uživatelům v srozumitelné podobě (tabulky, grafy). Pracuje prakticky pouze nad datovým skladem. Naproti tomu metody data miningu hledají v datech nové, dosud neznámé informace. Mohou pracovat nad daty z datového skladu, ale taky nad dalšími různými formáty dat.

8.1 Proces dobývání znalostí z databází

Na samotný proces dobývání znalostí z dat lze pohlížet ve dvou rovinách. Buďto z technického hlediska nebo managerského hlediska. Nejdříve je popsán managerský pohled (*Obrázek č. 21*). Impulsem pro zahájení samotného procesu dobývání znalostí bývá nějaký reálný problém. Cílem celého procesu pak je získání co možná nejvíce relevantních informací, které mohou přispět k řešení daného problému. Reálný problém se bude lišit podle organizace, která daný proces implementuje, může se jednat o nalezení určité skupiny zákazníků, která se vyznačuje specifikovanými vlastnostmi nebo může jít o klienty banky, kterým má být nabídnuta nová služba.

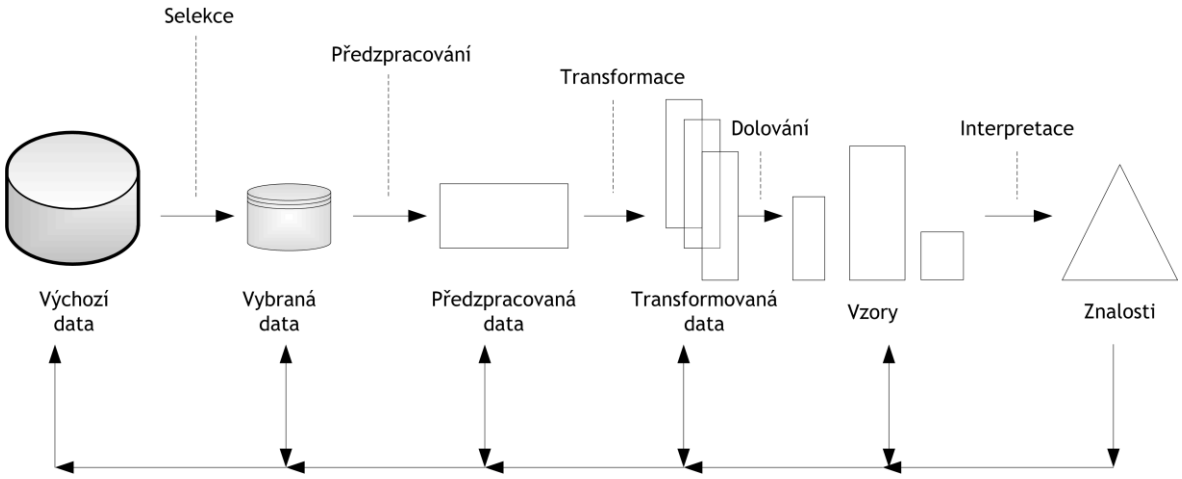


Obrázek č. 21: Managerský pohled na proces dobývání znalostí z databází [18]

Při řešení problému je nejprve nutné stanovit *řešitelský tým*. Jeho členové by měli být experti na danou problematiku, expert na data a expert na metody KDD.
Sestavený tým musí na začátku *specifikovat problém*, který je nutné řešit z pohledu dobývání znalostí. Jakmile je problém specifikován, je nutné získat veškerá *dostupná data*, která mohou být použita pro řešení problému. Je tedy nutné posoudit, zda dostupná data jsou relevantní k řešení daného problému. V některých případech je dokonce možné, že bude potřeba pracovat s daty, která jsou mimo databázi ve formě různých archivů nebo v nekonzistentních systémech. Náročnost získání a posouzení těchto poté rapidně stoupá. Cílem další etapy je zvolit vhodné *metody analýzy dat*. V rámci dobývání znalostí z databází se používá řada typů metod analýzy dat. Každá z nich je vhodná pro jiné účely a jiné typy dat. V mnoha případech je také nutné kombinovat více různých metod. Metod existuje celá řada například různé klasifikační metody, metody pro získávání asociačních pravidel, rozhodovací stromy, shlukovací algoritmy atd. Některé z těchto metod budou dále blíže popsány.

Během fáze *předzpracování dat* se získaná data pro řešení specifikovaného problému upravují do takové formy, jakou požaduje daná metoda pro dolování znalostí a také použitý nástroj. V této etapě může docházet také k doplnění chybějících hodnot, případně odstranění odlehlých pozorování.
Etapa *data mining* obsahuje aplikaci vybraných metod pro dolování znalostí. Zpravidla bývají metody aplikovány vícekrát za sebou, kdy jsou jednotlivé parametry upravovány na základě výsledků dosažených v minulém běhu. Takovýmto způsobem je dosaženo co možná nejlepších výsledků. Závěrečná etapa *interpretace* zahrnuje nezbytné zpracování obvykle velkého množství výstupů jednotlivých metod. Často se stává, že většina těchto výsledků obsahuje informace, které jsou z hlediska uživatele nezajímavé nebo samozřejmé. Je proto nutné odhalit ty opravdu užitečné výsledky, některé z nich lze použít ihned, další je nutné upravit do formy srozumitelné uživateli.

Naproti tomu z technického pohledu (*Obrázek č. 22*) je proces dolování znalostí z dat chápán jako interaktivní a iterativní proces, který tvoří kroky selekce, předzpracování, transformace, vlastního dolování (data mining) a interpretace [19]. Na rozdíl od prostého použití statistických metod se při procesu dolování znalostí z dat klade důraz na přípravu dat a vhodnou interpretaci výsledků. Při interpretaci se nalezené výsledky hodnotí z pohledu koncového uživatele.



Obrázek č. 22: Technický pohled na dobývání znalostí z databází

8.2 Předzpracování dat

Podobně jako v případě datových skladů, také data mining pro své metody potřebuje kvalitní a relevantní data. Tato etapa tak může být velice náročná, jelikož však na jejím pečlivém a kvalitním provedení závisí výsledek samotného dolování, měla by být této fázi věnována mimořádná pozornost. Mimo jiné je potřeba provést výběr atributů, které s řešeným problémem souvisí a jsou tedy z hlediska řešení daného problému relevantní. Dále je potřeba navrhnout, jakým způsobem se bude pracovat s chybějícími údaji – zda například budou doplněny určitou implicitní hodnotou nebo nebudou vůbec brány v potaz atd.

Dalším problémem, který se v této fázi řeší, je transformace dat. Některé data miningové algoritmy totiž pro svůj běh vyžadují data v určitém formátu – například pouze diskrétní nebo binární atd. Proto je potřeba mimo jiné spojité hodnoty vhodně diskretizovat. Tato kapitola si klade za cíl přiblížit etapu předzpracování dat pro potřeby dolování znalostí. Kapitola rovněž zahrnuje rozdělení dat z hlediska DM.

8.2.1 Dělení dat z hlediska dolování znalostí

Existuje mnoho způsobů, jak se dají data dělit. Každý databázový systém například obsahuje sadu datových typů, které dělí data podle jejich obsahu (datumové údaje, číselné, textové atd.). Pro potřeby data miningu lze data rozdělit na [20]:

- **Binární** – nabývají pouze dvou hodnot:
  Příklad:(0,1), (ano, ne), (muž, žena)

- **Kategoriální** – nabývají malého a konečného počtu hodnot a značí příslušnost k určité kategorii, udané svým očíslováním (0, 1, …, k) bez kvantitativního významu
  Příklad: národnost (česká, slovenská, polská…), očíslováno= 1: česká, 2:slovenská atd.

- **Ordinální** - nabývají také hodnot (0, 1,…, k), ale navíc je mezi nimi definováno uspořádání, případně bez významu vzdálenosti mezihodnotami. Obvykle se s nimi pracuje jako s kategoriálními daty.
  Příklad: známky ve škole (1-5) jsou uspořádány

- **Reálné (spojité)** – nabývají reálných hodnot z určitého intervalu
  Příklad: věk, výška, váha, cena atd.

8.2.2 Filtrace dat

Pod pojmem filtrace dat si lze představit řadu kroků, jejichž cílem je získání relevantních dat, která jsou připravena pro použití v dolovacích algoritmech. Filtrace mimo jiné zahrnuje:

- Výběr relevantních atributů pro analýzy
- Ošetření dat chybných, irelevantních, konstantních či redundantních
- Sjednocení formátů, měrných jednotek
- Transformace dat-kategorizace reálných hodnot, numerické zakódování

Při detailnějším pohledu na jednotlivé oblasti, které tato etapa zahrnuje, je patrné, že jde z větší části o podobné či stejné problémy, které je nutné řešit také v případě budování datového skladu, konkrétně pak ve fázi ETL. Tato problematika z pohledu datových skladů je detailně popsána v části 6.4.2. Problém transformace dat dnes zpravidla řeší použitý nástroj, kdy pomocí různých parametrů lze ovlivnit výslednou kategorizaci.

8.3 Algoritmy pro dolování znalostí z dat

Existuje celá řada algoritmů, které se používají pro data mining. Pro různá zadání a problémy jsou vhodné jiné algoritmy. Určitý algoritmus se použije v případě různých klasifikačních problémů, kdy například hledáme odpověď na otázku, zda bude reklamní kampaň úspěšná. Jiný algoritmus použijeme v případě určité segmentace, kdy budeme chtít rozdělit třeba zákazníky do určitých skupin podle zadaných kritérií. V této části budou popsány jedny z nejpoužívanějších algoritmů pro dolování znalostí z dat, které byly následně také prakticky aplikovány na testovaná data.

Data miningové metody z části také vychází ze statistiky a hojně využívají nejrůznějších statistických metod a poznatků, které dále rozšiřují. Jednou z hojně využívaných statistických veličin je **korelace**.

Korelace je míra závislosti mezi dvěma proměnnými. V některých případech může být korelace jednoduše vysvětlitelná, například společné nákupy mouky a cukru jsou motivované vysokým nákupem uvedeného zboží v případě, že máme k dispozici auto. V jiných případech však korelace nemusí být na první pohled jasná a zřejmá. Korelace může být pozitivní a negativní. Pozitivní korelace udává, že vysoká úroveň jedné proměnné bude provázena vysokou úrovní korelační proměnné. Naopak negativní korelace udává, že vysoká úroveň jedné proměnné bude provázena nízkou úrovní korelační proměnné. Jinými slovy čím více se zvětší hodnoty v první skupině, tím více se zmenší hodnoty v druhé skupině. Korelační koeficient (proměnná) nabývá hodnot od -1 do 1. Hodnota -1 značí negativní korelaci, hodnota 1 pozitivní. Pokud korelační koeficient je roven nule nebo se nule blíží, znamená to, že mezi sledovanými proměnnými není žádná statisticky zjistitelná lineární závislost. I v takovém případě však spolu proměnné mohou souviset, jen tento vztah nelze vyjádřit lineární funkcí.

8.3.1 Asociační pravidla

Asociační pravidla si lze zjednodušeně představit jako konstrukci IF-THEN, kterou lze nalézt v každém programovacím jazyce. Navíc se používají i v běžné mluvě. Asociační pravidla společně s rozhodovacími stromy patří mezi nejčastěji používané prostředky pro reprezentaci znalostí. Termín *asociační pravidla* byl široce zpopularizován na počátku 90. let minulého století v souvislosti s analýzou nákupního košíku. Při této analýze se zjišťuje, jaké druhy zboží si zákazníci současně kupují v supermarketech (například pivo a párek). Jde o hledání vazeb (asociací) mezi různým zbožím. Asociace lze také zaměnit za pojem analýza příčin a následků.

U pravidel, která jsou vytvořena z dat, nás obvykle zajímá, kolik příkladů splňuje předpoklad a kolik závěr pravidla, kolik příkladů splňuje předpoklad i závěr současně, kolik příkladů splňuje předpoklad a nesplňuje závěr atd. Jde tedy o pravidlo ve tvaru:

$$Ant \Rightarrow Suc,$$

kde *Ant* (předpoklad, příčina, levá strana pravidla, antecedent) i *Suc* (závěr, následek, pravá strana pravidla, sukcedent) jsou kombinace kategorií atributů. Základem pro asociační pravidla, je tzv. čtyřpolní tabulka. Její obecná podoba je zachycena v *Tabulce č. 8:*

	Suc	¬Suc	Σ
Ant	a	b	r
¬Ant	c	d	s
Σ	k	l	n

Tabulka č. 8: Ukázka tzv. čtyřpolní tabulky

Kde: $n(Ant \wedge Suc) = a$ je počet objektů pokrytých současně předpokladem i závěrem,
$n(Ant \wedge \neg Suc) = b$ je počet objektů pokrytých předpokladem a nepokrytých závěrem,
$n(\neg Ant \wedge Suc) = c$ je počet příkladů nepokrytých předpokladem ale pokrytých závěrem
$n(\neg Ant \wedge \neg Suc) = d$ je počet příkladů nepokrytých ani předpokladem ani závěrem.

$n(Ant)=a+b=r$, $n(\neg Ant)=c+d=s$, $n(Suc)=a+c=k$, $n(\neg Suc)=b+d=l$, $n=a+b+c+d$

Z těchto čísel (místo počet objektů se také používá termín *četnost*) lze počítat různé charakteristiky pravidel a hodnotit tak nalezené znalosti.

Mezi základní charakteristiky asociačních pravidel patří *podpora (support)* a *spolehlivost (confidence)*. Pomocí těchto vlastností lze ovlivnit běh samotného dolovacího algoritmu. Podpora je absolutní respektive relativní počet objektů, které splňují předpoklad i závěr. Je to tedy hodnota *a*, respektive:

$$P\,(Ant\ \wedge Suc) = \frac{a}{a+b+c+d}$$

Vzorec č. 1: Relativní vyjádření podpory

Spolehlivost, také nazývaná platnost je vlastně podmíněná pravděpodobnost závěru pokud platí předpoklad, tedy:

$$P\,(Ant \mid Suc) = \frac{a}{a+b}$$

Vzorec č. 2: Vyjádření spolehlivosti

Běh algoritmu se poté zpravidla ovlivňuje nastavováním parametrů nejčastěji nazvaných minimální podpora *(min support*) a minimální spolehlivost (*min confidence*), kdy pomocí těchto parametrů stanovujeme minimální hodnoty, kterých je potřeba dosáhnout, aby bylo pravidlo vůbec vygenerováno. V případě podpory lze hodnotu stanovovat, jak absolutně, tak relativně. Záleží na použitém nástroji (pokud takové vyjádření podporuje). Pravidlo vygenerované algoritmem pak například může vypadat následovně:

Pivo ∧ Párek ⇒ Hořčice s podporou 10.

Algoritmy pro hledání asociací

Algoritmy pro dolování asociací se snaží automatizovat proces provádění statistických testů. Tyto algoritmy automaticky generují a testují všechny možné hypotézy a na výstup generují ty, které splňují zadaná kritéria (hodnoty minimální spolehlivost a minimální podpory). Hlavním problémem těchto algoritmů pro hledání asociací je časová složitost, protože testují všechny možné sentence. Počet generovaných (a testovaných) kombinací je exponenciálně závislý na počtu atributů [21]. Tedy v případě většího počtu atributů, významně rostou požadavky na čas, potřebný k nalezení všech možných asociačních pravidel. Řešení spočívá v nalezení rychlých algoritmů, které jsou schopny nalézt všechny požadované hypotézy, ale omezit počet testů.
Algoritmy pro generování asociačních pravidel lze dělit na [20]:

- **Triviální** – generují a testují všechny možné sentence jako kombinace hodnot atributů, délek antecedentů a sukcedentu, kombinací dvojic antecedentu a sukcedentu. Pro větší data jsou tyto algoritmy nepoužitelné $\Rightarrow$ vysoká časová složitost.

- **Uspořádané generování pravidel** – vhodně uspořádané postupné prodlužování délky antecedentu a sukcedentu. Pokud *a* (hodnota z čtyřpolní tabulky) < *minimální podpora,* již se negenerují další pravidla s větší délkou.

- **Vzorkování** – rozsáhlá data se zpracovávají po částech, hledají se vhodní kandidáti pro hypotézy, následně se přes celá data testují pouze tito kandidáti.

Algoritmus Apriori

Patří mezi jedny z nejpoužívanější asociačních algoritmů. Jeho implementaci lze nalézt ve velkém počtu nástrojů pro DM. Algoritmus navrhl R. Agrawal v souvislosti s analýzou nákupního košíku. Jádrem algoritmu je hledání často se opakujících množin položek (*frequent itemsets*). Jde tedy o konjunkce kategorií (hodnot atributů), které dosahují předem zadané četnosti (*minimální podpora*) v datech.

Při hledání kombinací délky *k*, které mají vysokou četnost, se využívá toho, že již je známa kombinace délky k-1. Proto při vytváření kombinace délky k, se spojují kombinace délky *k-1*. Při vytváření jedné kombinace délky *k,* musí být opět splněn požadavek minimální četnosti (podpory) u všech podkombinací *k-1*. Například z tříčlenných kombinací {A_1,A_2,A_3 A_1,A_2,A_4 A_1,A_3,A_4 A_2,A_3,A_4}, které dosahují minimální podpory lze vytvořit pouze jedinou čtyřčlennou kombinaci A_1,A_2, A_3,A_4.

8.3.2 Rozhodovací stromy

Hledají podmínky ve formě hodnot vstupních atributů, na jejich základě je pak objekt zařazen do příslušné klasifikační třídy. Název rozhodovací se používá proto, jelikož se výsledná pravidla přehledně zobrazují ve formě stromu. Uzly zobrazují atributy, podle nichž se právě dělí, hrany reprezentují hodnoty atributů. Na rozdíl od asociací RS nehledají, nýbrž konstruují optimální implikace mezi množinou vstupních atributů a předpovídaným atributem. Tedy jde metodu pro získávání znalostí z dat, která v datech hledá charakteristický popis zadaných tříd pomocí kombinací hodnot atributů [20].

Klíčovou otázkou pro běh samotného algoritmu je, jak vybrat vhodný atribut pro větvení stromu. Přičemž cílem je vybrat takový atribut, který od sebe nejlépe odliší příklady různých tříd. Vodítkem v tomto smyslu jsou pro algoritmus charakteristiky atributu převzaté z teorie informace nebo pravděpodobnosti: entropie, informační zisk, Gini index a tak dále. Nejčastěji se u jednotlivých nástrojů a jejich implementace konkrétního algoritmu používá Entropie.

Entropie je pojem používaný v přírodních vědách a vyjadřuje míru neuspořádanosti nějakého systému. V teorii informace je entropie definována jako funkce [21]

$$H = -\sum_{t=1}^{T}(p_t * log_2\, p_t),$$

Vzorec č. 3: Definice entropie

kde p_t je pravděpodobnost výskytu třídy t (relativní četnost třídy t počítaná na určité množině příkladů) a T je počet tříd.

Výpočet entropie pro jeden atribut se provádí následujícím způsobem:

1. Pro každou hodnotu *v*, kterou může nabýt uvažovaný atribut *A* se spočte entropie *H(A(v))* na skupině příkladů, které jsou pokryty kategorií *A(v)*

$$H\big(A(v)\big) = -\sum_{t=1}^{T}\frac{n_t(A(v))}{n(A(v))}\, log\, \frac{n_t(A(v))}{n(A(v))}$$

Vzorec č. 4: Výpočet entropie pro jeden atribut

2. Se spočte střední entropie *H(A)* jako vážený součet entropií *H(A(v))*, přičemž váha v součtu jsou relativní četnosti kategorií *A(v)*

$$H(A) = \sum_{v\,\in Val\,(A)}\frac{n(A(v))}{n}\, H(A(v))$$

Vzorec č. 5: Výpočet střední entropie

Pro větvení stromu se pak použije atribut s nejmenší entropií *H(A)*.

Algoritmů pro konstrukci rozhodovacích stromů existuje řada. Většina z nich je však variantou základního algoritmu. Tento postup bývá často nazýván top *down indruction of decision trees* (TDIDT). Využívá se zde metody „rozděl a panuj“ (*divide and concquer*). Trénovací data se postupně rozdělují na menší a menší podmnožiny (uzly stromu) tak, aby v těchto podmnožinách převládaly příklady jedné třídy.

Podrobný popis algoritmu, ale i teorie související s rozhodovacími stromy lze nalézt například zde [20] nebo zde [21].

8.3.3 Shlukování

Shlukování se snaží ověřit, zda se data (množina objektů) rozpadají na nějaké výrazné podmnožiny – objekty vzájemně si podobné (shluky). Pokud ano, měli bychom být schopni tyto shluky charakterizovat – definovat jejich typické vlastnosti. Čili shlukování se snaží rozdělit data do určitých podmnožin na základě jejich vlastností (hodnot atributů).

Shlukování sebou nese celou řadu problémů. Mimo jiné neexistuje jednoznačná definice podobnosti objektů a rovněž neexistuje ani jednoznačná definice shluku. Nejedná se tedy o ucelenou teorii, nýbrž jde o řadu metod, které jsou založeny na různých principech. Mezi základní problémy, které vznikají při shlukování, například patří:

- Výběr atributů charakterizujících podobnost
- Podobnost respektive nepodobnost (vzdálenost objektů)
- Pojem vzdálenosti shluků
- Počet shluků rozkladu
- a jiné

Vzhledem k rozsáhlosti problematiky a zaměření práce zde nebude tato oblast detailně diskutována. Naznačeny budou pouze základní principy. Detailnější popis lze nalézt například v [20].

Míra vzdálenosti objektů

Udává, jak jsou od sebe objekty vzdálené. Jinými slovy říká, jak jsou si dané objekty podobné respektive nepodobné. Pro měření vzdálenosti objektů se používají metriky. Metriky vycházejí z geometrického modelu dat, kde objekty o *n* znacích chápeme jako body v *n-rozměrném Euklidovském prostoru* E_n. Pak lze podobnost objektů vyjadřovat jako vzdálenost odpovídajících bodů. Metrika *V* je funkce, která přiřazuje každé dvojici bodů (Oi, Oj) číslo *V* takové, že platí:

$V(Oi, Oj) = 0 \Leftrightarrow Oi = Oj$
$V(Oi, Oj) \geq 0$
$V(Oi, Oj) = V(Oj, Oi)$
$V(Oi, Oj) + V(Oj, Ok) \geq V(Oi, Ok)$

Nejznámější metrikou je tzv. Eukleidovská vzdálenost:

$$V(Oi, Oj) = \sqrt{\sum_{i=1}^{n} (ai - bi)^2}$$

Vzorec č. 6: Eukleidovská vzdálenost

kde *ai, bi* jsou jednotlivé hodnoty objektů *Oi* respektive *Oj*.

Co je to shluk

Existuje řada popisů chápání shluku. Jednoznačná definice však není stanovena a proto lze v literatuře nalézt definic mnoho. Jedna z nich definuje shluk takto:

Je dána množina objektů O = {O1, ..., Om} a míra vzdálenosti objektů V. Shlukem nazveme takovou podmnožinu X ∈ O, pro niž platí

$$max_{Oi,Oj \in X} V(Oi, Oj) < min_{Oi \in X,\ Ok \notin X}(Ok, Oi)$$

Typy shlukovacích metod

Tak jako existuje mnoho definic shluku, tak také existuje celá řada metod (algoritmů), které se snaží řešit problém shlukování. Rozdělit je lze do několika skupin:

Podle cíle shlukování

- *nehierarchické* – výsledkem je prostý rozklad objektů na podmnožiny
- *hierarchické* – produkují hierarchii rozkladu, kde každý shluk je zjemněním předcházejícího

Podle tvaru výsledných shluků

- *kulové shluky* – body soustředěné pravidelně kolem těžiště shluku
- *obecné shluky* – tvoří souvislé husté oblasti nejrůznější tvarů

a další skupiny metod, kdy lze metody rozdělit například podle použitého typu algoritmu (sekvenční, heuristické, rekurzivní atd.).

K-středová metoda

Patří mezi nejpoužívanější shlukovací metodu a vyskytuje se ve většině nástrojů pro analýzy dat. Jde o nehierarchickou optimalizační metodu, čili se snaží najít nejlepší prostý rozklad mn*ožiny objektů iteračním způsobem. Základní verze algoritmu má následující kroky:*

1. *zadání **k** počátečních bodů*
2. *přiřazení každého bodu k nejbližšímu typickému bodu a jemu odpovídajícímu shluku*
3. *výpočet těžiště každého **z** k shluků*
4. *definování nový typických bodů ve vypočtených těžištích*
5. *pokud došlo ke změně v přiřazení bodů shlukům, opakování od bodu 2*
6. *výpočet charakteristik výsledného rozkladu*

9 Vlastní řešení data miningu

Tato kapitola popisuje aplikaci dříve popsaných principů a metod na vlastních datech. Již během OLAP analýzy byly odhaleny zajímavé hypotézy, které byly v této fázi ověřovány. Kromě nich byla provedena celá řada dalších testů za účelem nalézt v datech další zajímavé informace. Cílem této kapitoly je popsat průběh dolování, použité nástroje a také získané výsledky a možnost jejich dalšího využití.

Pro dolování znalostí byly použity dva nástroje. *Business Intelligence Development Studio* – podobně jako v případě OLAP analýzy. Tento nástroj umožňuje v dobře známém prostředí provádět také dolování znalostí z dat a to prakticky pomocí stejných principů, aplikovaných u datových skladů. Kromě toho, byl pro některé analýzy použit také open source nástroj určený pro data mining. Konkrétně pak *RapidMiner*, který v kategorii zdarma dostupných nástrojů patří mezi velice oblíbené. Další nástroj byl použit zejména z důvodu porovnání obou nástrojů, které lze v kapitole také nalézt. A to porovnání, jak dosažených výsledků, tak způsob práce s aplikací.

9.1 Předzpracování dat

Průběh této etapy byl zjednodušen, jelikož důkladná analýza dat a předzpracování bylo provedeno již při budování datového skladu. Detailně je tato problematika popsána v sedmé kapitole. Pro potřeby data miningu byly pouze ze zdrojových tabulek DS vytvořeny pohledy, které obsahují pouze relevantní data určená pro samotné dolování. Tento krok je důležitý z pohledu samotných dolovacích algoritmů, jelikož to, že budou pracovat pouze s relevantními daty, může značně zrychlit jejich běh. Celkem byly vytvořeny 3 pohledy:

- **DT_asociace_shlukování** s atributy:
  - *Aukce_je_kup_ted, aukce_status, kvartal, mesic, den_nazev, obor, oblast, název_kategorie, prodejce_login, zákazník_login, vyvolávací_cena, prodejní_cena, kup_teď_cena, počet_příhozů, fakt_id*

- **DT_prémiové_kategorie** s atributy:
  - *Obor, oblast, vyvolávací_cena, kup_teď_cena, fakt_id*

- **DT_aukce_status** *s atributy:*
  - *Obor, oblast, vyvolávací_cena, kup_teď_cena, aukce_status, aukce_id*

Další úkoly spojené s předzpracováním dat, jako je například diskretizace spojitých atributů pro potřeby jednotlivých algoritmů, již řeší samotné data miningové nástroje a problematika je zmíněna přímo u popisu průběhu dolování.

9.2 Rozhodovací stromy

Pomocí rozhodovacích stromů jsem se snažil ověřit dvě hypotézy:

1. **Má na úspěšný/neúspěšný prodej zboží vliv vyvolávací cena zboží a její kategorie?**
2. **Existují v systému prémiové kategorie? Čili ovlivňuje vyvolávací cena a kup teď cena kategorii, ve které je aukce umístěna?**

Kapitola je koncipována tak, že nejprve je popsáno dolování s využitím nástroje BIDS. Poté následuje srovnání s RapidMinerem. Na závěr je pak uvedeno zhodnocení výsledků, respektive možnost jejich využití v obchodním prostředí.

9.2.1 Dolování pomocí BIDS

Pro potřeby prvního rozhodovacího stromu byl použit pohled *DT_aukce_status*, který obsahuje potřebné atributy pro tuto analýzu.

Vstupní atributy:

- Vyvolávací cena
- Kup teď cena
- Kategorie zboží

Předpovídaný atribut:

- Aukce status

Již z výsledků, získaných z analýzy datového skladu je známo, že převažuje zboží, které se neprodá (62 %). Nyní však můžeme získat informaci, zda existují přímo kategorie, ve kterých je prodej zboží náročnější než v jiných kategorií a také zda počáteční cena zboží ovlivňuje to, zda daná aukce skončí prodejem či nikoliv. Na druhou stranu můžeme objevit i kategorii, ve které se naopak zboží prodává více než dobře.

Nastavení algoritmu

Microsoft SQL Server 2005 nabízí pro rozhodovací stromy algoritmus, který je nazván *Microsoft Decision Trees Algorithm*. Ten umí pracovat jak s diskrétními, tak spojitými hodnotami, přičemž na základě typu proměnných se liší výpočet algoritmu. Proto aby algoritmus mohl správně pracovat, vyžaduje jeden klíčový atribut, vstupní atributy a jeden předpovídaný atribut. Nastavení parametrů algoritmu a vstupních hodnot je velice jednoduché a prakticky totožné s jakoukoliv jinou činností prováděnou v rámci BIDS. Ke všemu slouží přehlední průvodci, pomocí kterých je celý proces definován. Následné nastavení parametrů algoritmu se provádí pomocí karty vlastnosti, podobně jako například v případě ovládacích prvků grafického uživatelského rozhraní.

Algoritmus nabízí několik parametrů, pomocí kterých lze ovlivnit průběh dolování a jeho výsledek. S jejich nastavením bylo experimentováno, aby dosaženo co možná nejlepších výsledků. Parametry, kterým byly upravovány hodnoty oproti výchozímu nastavení, jsou:

- MINIMUM_SUPPORT
- SCORE_METHOD

Ostatní parametry byly nakonec ponechány ve výchozím nastavení, jelikož jejich změna neovlivňovala dosažený výsledek. Detailní popis algoritmu včetně jednotlivých parametrů lze nalézt zde [12].

Dále byly rovněž zkoušeny veškeré metody (na převod spojitých atributů na diskrétní), které *Bussiness Inteligence Development Studio* nabízí. Jako správná volba se nakonec projevila možnost *clustered*, která ve výsledku vytvořila intervaly, více vyhovující povaze dat (pro každý uzel, rozdílné intervaly). Druhá možnost – *EqualAreas* totiž tvoří intervaly, které musí obsahovat stejný počet hodnot a všechny uzly stromu musí použít stejné intervaly (např. 0-100, 101-1000, 1001 a výš). Navíc povaha zdrojových dat rovněž není pro tuto metodu vhodná. Například pokud bychom vzali v potaz prémiovou kategorii Zlaté mince, kde se vyvolávací ceny mohou pohybovat řádově o sto či tisícikoruny výš než v případě běžného zboží, je pro nás nevýhodné mít stanoveny pro všechny kategorie stejné intervaly. Může totiž poté docházet ke zkreslení a špatné interpretaci výsledků.

BIDS rovněž nabízí možnost automaticky zvolit metodu pro převod spojitých atributů na diskrétní. Pokud byla vybrána tato varianta, BIDS správně vybralo metodu *clustered*. Počet vytvářených intervalů byl nastaven na čtyři. Tento počet byl zvolen jako kompromis. V některých případech by totiž jemnější dělení přineslo detailnější rozložení zboží, co se týče jejich vyvolávacích cen. Ovšem v některých kategoriích není velký rozptyl ve vyvolávacích cenách a více intervalů zbytečně více zjemňovalo dělení. Toto jemnější dělení už pak nemělo žádný informační zisk, například vznikaly intervaly s vyvolávací cenou 1-15 Kč, 16-30 Kč apod. Proto se počet čtyř intervalů jevil jako optimální pro obě možnosti.

Průběh dolování

Jak je popsáno v části věnované nastavení algoritmu, proces sestavení rozhodovacího stromu byl nesčetně krát opakován s nejrůznějším nastavením zmíněných parametrů, dokud nebylo dosaženo optimálních výsledků.

Parametr SCORE_METHOD byl nastaven na hodnoty: *Entropy a Bayesian Dirichlet Equivalent with Uniform Prior*. V obou případech bylo dosaženo stejných výsledků, proto byl nakonec tento atribut ponechán ve výchozím nastavení (*Bayesian Dirichlet Equivalent with Uniform Prior*).

Parametr MINIMUM_SUPPORT byl volen ve velice širokém rozpětí hodnot – od 10 až po 12 000:

- V případě nízkých hodnot 10 – 50 bylo dosaženo detailnějších výsledků, které jsou popsány v části výsledky. Rozhodovací strom měl v tomto případě pouze 3 – 4 úrovně, takže byl také snáze čitelný. Strom rovněž obsahoval veškeré kategorie hned v první úrovni, kdy v dalších úrovních stromu byly intervaly vyvolávací ceny a kup teď ceny.

- V případě nastavení parametru na vyšší hodnoty 100 – 12 000 dochází ke značné restrukturalizaci stromu. Ten nyní na nejvyšší úrovni obsahuje pouze 2 hodnoty. A to zda aukce pochází z kategorie Asie (nejméně úspěšná) nebo nepochází. Stejná situace se opakuje v dalších úrovních. Konstrukce postupuje od těch nejméně úspěšných kategorií ve vrchních úrovních stromu směrem dolů k úspěšnějším. Počet úrovní je značně vyšší, pohybuje se v rozmezí 10 – 20 úrovní podle nastavení minimální podpory. Čím větší číslo, tím méně úrovní, jelikož se uzly s menším počtem výskytů ztrácejí. Čitelnost stromu je v tomto případě značně obtížnější. Nicméně lze se dopátrat stejných výsledků jako výše pokud hodnota minimální podpory není nastavena na příliš vysokou hodnotu.

Jako optimální je tedy vhodné navolit parametr minimální podpory na nízkou hodnotu v intervalu 10 – 50.

Výsledky

Ve výsledném stromu lze nalézt rozličné uzly. Nejčastěji se vyskytují uzly, kde je pravděpodobnost prodeje či „neprodeje“ zboží velice vyrovnaná. V některých uzlech je pravděpodobnější prodej, v jiných naopak neprodej. Pokud bychom chtěli určit pravděpodobnost, zda se zboží prodá pouze na základě toho, v jaké kategorii je umístěno, pak lze ve stromu identifikovat několik kategorií, u kterých je pravděpodobnost prodeje aspoň 50%:

- Středověk (53,9 %)
- ČR do roku 1939 (59,7 %)
- Antické (51,7 %)
- ČR po roce 1939 (52,4 %)

Pokud tedy budeme obchodovat se zbožím z těchto kategorií, již máme 50% šanci na úspěšný prodej. Zajímavější výsledky však získáme, pokud kategorii doplníme ještě o vyvolávací cenu zboží (další úroveň stromu). Pravděpodobnost prodeje se zvýší, tento nárůst však nelze považovat za významný. Prakticky pouze definuje interval pro vyvolávací cenu, ve kterém jsou prodeje nejúspěšnější:

- Středověk a vyvolávací cena < 1300 Kč (59,4 %)
- ČR do roku 1939 a vyvolávací cena < 1300 Kč (62,4 %)
- Antické a vyvolávací cena < 1300 Kč (55,6 %)
- ČR po roce 1939 a vyvolávací cena < 180 Kč (53,5 %)

V této úrovni stromu lze najít i další zajímavé uzly, které jsme na vyšší úrovni stromu nepovažovali za významné. Například pokud je aukce z kategorie zlaté mince a její vyvolávací cena je menší jak 1300 Kč, pravděpodobnost prodeje je 62,2 %. Nebo jiný příklad se stejnou kategorií, pokud byla cena kup teď stanovena na hodnotu menší jak 180 Kč, tak pravděpodobnost prodeje byla dokonce 85,4%.

Z daného rozhodovacího stromu však jsme schopni získat i opačnou informaci – tzn., které kategorie patří mezi ty nejméně úspěšné, co se týče pravděpodobnosti prodeje zboží (*Tabulka č. 9):*

Název kategorie	Pravděpodobnost neprodeje v %
Asie	71,4
Afrika	70,8
Amerika	68,8
Evropa	63,1
Austrálie a Oceánie	62,5
Euro Mince	62,4
Medaile	60,3
Severní Ameerika	57

Tabulka č. 9: Nejméně úspěšné kategorie

Závěr

Rozhodovací strom ve zdrojové datové struktuře skutečně odhalil kategorie, jejichž zboží má vyšší pravděpodobnost prodeje pouze na základě umístění v této kategorii než zboží v jiné kategorii.

Lze tedy říci, že na úspěšný prodej zboží má vliv to, v jaké kategorii bude umístěno a také jaká bude jeho vyvolávací cena.

Rovněž bylo odhaleno, v jakém rozmezí je nejvhodnější volit vyvolávací cenu, abychom pravděpodobnost prodeje ještě zvýšili. Dalším zajímavým výsledkem je naopak identifikace kategorií, ve kterých se zboží prodává velice těžko.

Rozhodovací strom č. 2

V tomto případě jsem se snažil ověřit zda v systému existují prémiové kategorie, to znamená, zda vyvolávací cena a kup teď cena má vliv na to, v jaké kategorii je zboží umístěno.

Vstupní atributy:

- vyvolávací cena
- kup teď cena

Předpovídaný atribut:

- kategorie

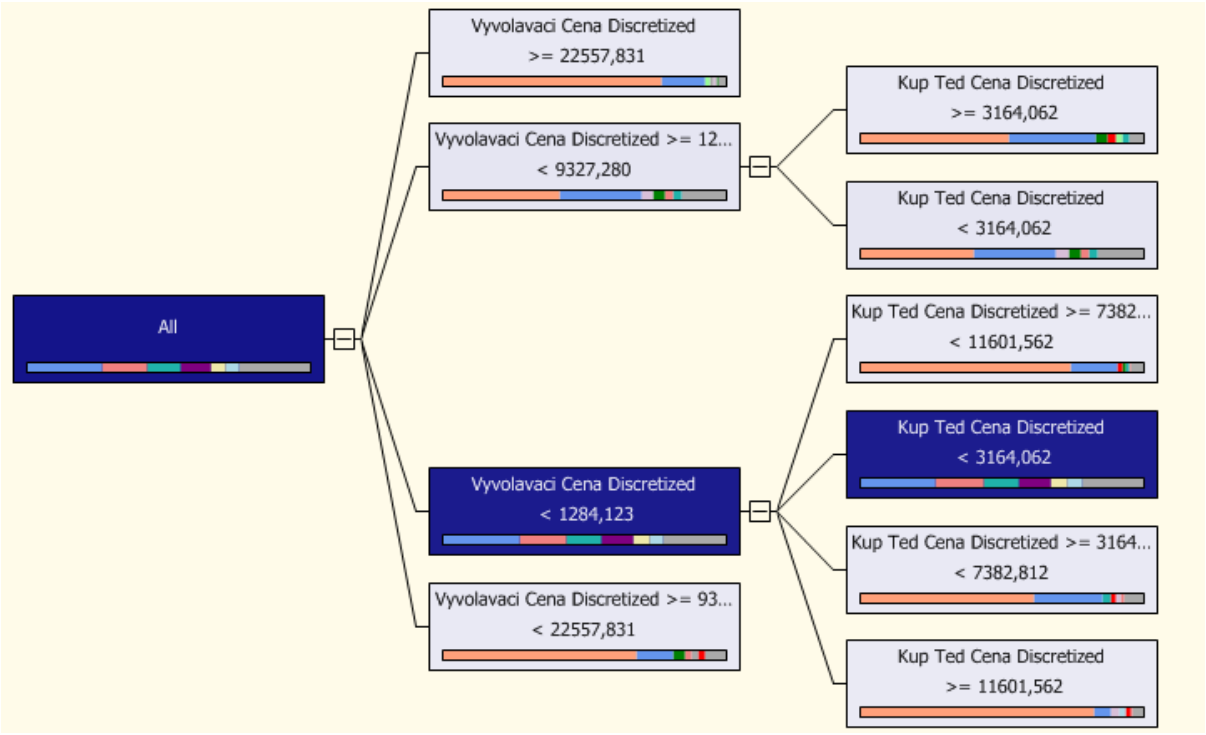
Jelikož vstupní atributy jsou spojité, bylo potřeba je opět vhodným způsobem převést na diskrétní. Poněvadž se jedná o stejné atributy jako v případě minulého rozhodovacího stromu, bylo využito již získaných poznatků (vytvořeny celkem čtyři intervaly).

Nastavení algoritmu

Podobně jako v předchozím případě bylo i zde experimentováno s nastavením parametrů algoritmu tak, aby bylo dosaženo co možná nejlepších výsledků.
Opět byly měněny pouze hodnoty parametrů SCORE_METHOD a MINIMAL_SUPPORT. U ostatních parametrů, vzhledem k testovacím datům, nebyl důvod měnit jejich výchozí hodnoty.

Průběh dolování

Parametr minimální podpory byl opět nastavován v širokém rozmezí od 10 až po hodnotu 2000. Pokud byl parametr nastaven na nižší hodnotu (10-1000), měl výsledný strom celkem 3 úrovně, kdy poslední úroveň stromu detailněji specifikuje kup teď cenu pro jednotlivé intervaly vyvolávací ceny. Struktura stromu je zachycena na *obrázku č 23*. Uvedený obrázek pochází přímo z prostředí BIDS, kde je přímo možné analyzovat hloubku jednotlivých závislostí a vztahy uvnitř nalezené hierarchie. Čím je barva políčka tmavší, tím je pravděpodobnost výskytu této vlastnosti vyšší. Každý obdélník diagramu (uzel stromu) rovněž obsahuje poměrový indikátor.



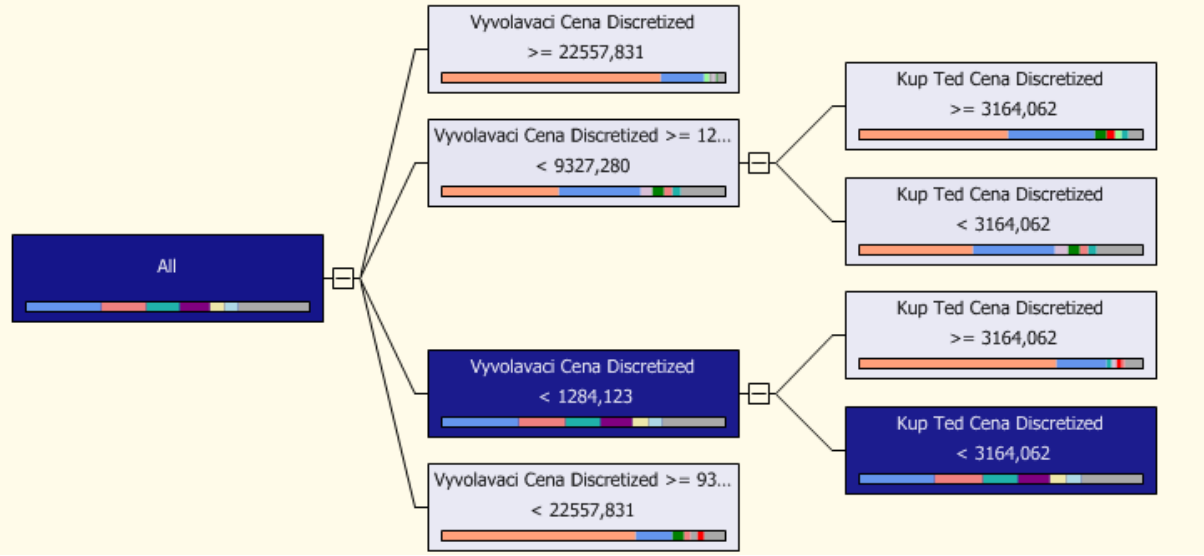
Obrázek č. 23: Rozhodovací strom Prémiové kategorie s parametrem minimální podpory na hodnotě 100

Legenda:

Zlaté mince
Česko

Pokud je parametr minimální podpory nastaven na vyšší hodnoty (1000 – 2000) dochází postupně k spojení některých uzlů na třetí úrovni stromu (*obrázek č 24.)*.

Takže ve výsledku je získán jednodušší strom, ovšem začínají se ztrácet některé podrobnější informace, které mohou být prospěšné. Pokud bychom pokračovali ve zvyšování hodnoty parametru minimální podpory, ve výsledku bychom získávali stále menší a detailů zbavený strom, který již nemá takovou informační hodnotu.



Obrázek 24: Rozhodovací strom Prémiové kategorie s parametrem minimální podpory na hodnotě 2000

Pro parametr SCORE_METHOD byly vyzkoušeny všechny možnosti, které je možné použít pro diskrétní hodnoty. Ve všech případech bylo dosaženo stejných výsledků, takže ve finále byl tento parametr ponechán ve výchozím nastavení.

Výsledek

Závěrem lze říci, že ve zdrojové databázi existují prémiové kategorie. Konkrétně jde o jednu kategorii, která se ve větším počtu vyskytuje prakticky ve všech větvích stromů, kde se vyvolávací cena zboží odchyluje od normálu. Jde o kategorii *Zlaté mince.*

Pokud je vyvolávací cena z intervalu od 9 300 do 22 600 Kč, pak s 65,8% pravděpodobností aukce s takovouto vyvolávací cenou pochází právě z kategorie Zlaté mince. Jestliže je vyvolávací cena větší jak 22 600 Kč, pak se 74,8 % pravděpodobností aukce opět pochází z kategorie *Zlaté mince*.

Pokud vezmeme v potaz ještě kup teď cenu (další úroveň stromu a poslední úroveň hierarchie) pak:

- Jestliže je vyvolávací cena menší jak 1300 Kč a zároveň je vyvolávací cena větší jak 11 600 Kč, pak s pravděpodobností 81,2 % je aukce umístěna v kategorii *Zlaté mince*.

Kromě zjištění této prémiové kategorie lze z rozhodovacího stromu vyčíst ještě jednu podstatnou informaci, a to s jakou vyvolávací cenou se můžeme setkat nejčastěji. 93,8 % všech aukcí má vyvolávací cenu nižší než 1300 Kč a 93 % ze všech aukcí má kup teď cenu nižší, jak 3200 Kč. Na obrázcích jsou tyto uzly vyznačeny tmavomodrou barvou (nejčastěji se vyskytující hodnoty).

9.2.2 Dolování pomocí RapidMineru

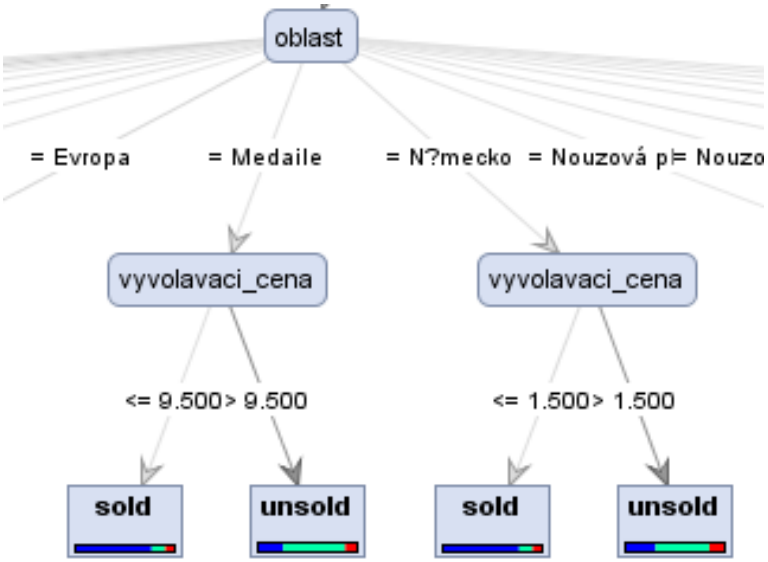
V případě Rapid Mineru bylo nejprve nutné zpřístupnit aplikaci data, nad kterými se bude pracovat. Data byla do aplikace importována z CSV souboru. Pro použití rozhodovacích stromů nebylo nutné příslušné atributy dále upravovat (*aukce_status, kup_teď_cena, vyvolávací_cena, kategorie*). Pouze bylo nutné označit předpovídaný atribut.

Nastavení algoritmu

Při nastavování algoritmu bylo využito zkušeností nabytých během dolování prostřednictvím BIDS. Přistoupil jsem proto ihned k nastavení minimální podpory-zde byl tento atribut nazván jako *minimal leaf size (*minimální velikost listu stromu). Kromě tohoto parametru RapidMiner nabízí ještě jeden podobný parametr a to *minimal size for split* – čili minimální velikost uzlu, která je potřebná k tomu, aby mohlo dojít rozdělení stromu. Tento parametr byl nastavován na stejnou hodnotu jako *minimal leaf size.* Jako kritérium pro sestavování stromu bylo použito *gini_index.* Při volbě ostatních možností software nevygeneroval žádný strom. Atribut maximální hloubky stromu (maximal_depth) byl nastaven na hodnotu 4. Hlubší strom pouze více komplikoval výsledky, navíc při ponechání výchozí hodnoty (20) docházelo k pádům aplikace.

Výsledky

Bylo dosaženo stejných výsledků jako v případě použití BIDS. RapidMiner sice zkonstruoval strom poněkud jinak, nicméně po detailnějším zkoumání byly výsledky totožné. Nevýhoda oproti BIDS je v tom, že tento software neposkytuje procentuální vyjádření jednotlivých hodnot předpovídaného atributu v uzlech, ale nabízí pouze absolutní čísla. Tedy interpretace výsledků je v tomto případě obtížnější, pokud jsou požadovány procentuální podíly.
Vygenerovaný strom (*obrázek č. 25*) má odlišnou strukturu než v případě BIDS. V tomto případě je kořenovým uzlem atribut *kup_teď_cena*, která má 2 větve: ≤ 4500 a > 4500, poté již se objevují jednotlivé kategorie a v listech poté hodnoty předpovídaného atributu (*sold, unsold*):



Obrázek č. 25: Ukázka výstupu (část stromu) z aplikace RapidMiner

Na rozdíl od BIDS, ale RapidMiner nabízí širší možnosti grafického zobrazení výsledků, kromě klasického stromu lze vybírat přibližně z 10 nejrůznější grafů (např. zobrazení v kruhu místo stromu). Navíc také nabízí textový výstup:

```
kup_ted_cena <= 4.500
|   oblast = CR do 1939
|   |   vyvolavaci_cena <= 100.500: sold {sold=5688, unsold=1117}
|   |   vyvolavaci_cena > 100.500: unsold {sold=2674, unsold=3354}
|   oblast = CR po 1939
|   |   vyvolavaci_cena <= 21.500: sold {sold=10962, unsold=1299}
|   |   vyvolavaci_cena > 21.500: unsold {sold=15068, unsold=16052}
|   oblast = Cesko
|   |   vyvolavaci_cena <= 4.500: sold {sold=17008, unsold=2327,
killed=79}
|   |   vyvolavaci_cena > 4.500: unsold {sold=24822, unsold=41306}
|   oblast = Afrika: unsold {sold=2716, unsold=11024}
```

Průběh dolování

Samozřejmě bylo experimentováno s nastavením jednotlivých parametrů algoritmu jednak za účelem ověření toho, že daný strom skutečně vykazuje stejné vlastnosti jako strom vytvořený v aplikaci BIDS a za další za účelem zda se náhodou neobjeví nové skutečnosti dříve neodhalené. Při zvyšování minimální podpory se strom skutečně choval jako v předchozím případě (ztráta detailů). Navíc při experimentování s parametry algoritmu občas docházelo k pádu aplikace z důvodu nedostatku paměti (díky velkému množství dat), což se v případě BIDS nikdy nestalo.

9.2.3 Zhodnocení výsledků

Pomocí rozhodovacích stromů byla zjištěna řada zajímavých informací. Tato část textu se snaží nastínit, jakým způsobem by se dosažené výsledky daly využít a interpretovat v obchodním prostředí. Ať už v případě analýzy trhu, kterou může provádět potencionální podnikatel, který se rozhodne podnikat v této oblasti nebo provozovatel aukčního portálu, který nabízí k prodeji také komodity z oblasti numismatiky a bankovek. Každý z nich může dosažené výsledky interpretovat a využít z jiného pohledu. Pro potencionálního podnikatele lze dosažené výsledky interpretovat v tomto smyslu:

Pokud se podnikatel chce orientovat na zboží s vyšší pořizovací cenou, a tedy i vyššími výdělky, měl by volit komodity z oblasti zlatých mincí. Nevýhodou na druhou stranu oproti vyšším výdělkům může být nižší podíl na trhu a menší cílová skupina zákazníků než u komodit, které se nabízejí s podstatně nižšími vyvolávacími cenami, ale pokrývají přes 90% trhu. Pro zajímavost zboží z kategorie Zlaté mince zabírá pouze 4,5 % trhu. Pokud budeme volit zboží s vyvolávací cenou do 1300 Kč, čili nejrozšířenější část trhu máme šanci oslovit početnější skupinu zákazníků napříč kategoriemi a oblastmi (bankovky, numismatika). Jestliže se jedná o začínající společnost bez patřičného zázemí a povědomí mezi zákazníky, je toto určitě nejlepší volba, protože pokud by se ihned ze začátku pokusila nabízet zboží z prémiové kategorie, nemusela by společnost uspět.

A to zejména z toho důvodu, že zákazníci orientující se na prémiové zboží, určitě dají přednost již ověřené a zaběhnuté společnosti (prodejci) s dobrým jménem. Někteří zákazníci již mají své oblíbené prodejce a přesvědčit je, aby začali nakupovat jinde, může být velice složité.

Další zjištění, které je vhodné zohlednit při strategických rozhodnutích:

- Zcela by se měli vyhnout obchodování s artikly z kategorií, kde pravděpodobnost jejich prodeje je velice nízká (Asie, Afrika)

- Orientovat by se tedy měli na populární zboží, kde pravděpodobnost prodeje patří mezi nejvyšší. Zde připadají v úvahu dvě varianty:

    - Zvolí zboží, které patří mezi ty nejpopulárnější a je ho na trhu nejvíce. Tím pádem připadá v úvahu větší množina potencionálních zákazníků, ale samozřejmě také větší konkurence.

    - Další možností je zvolit zboží s vysokou pravděpodobností prodeje, ale menší oblíbeností. Zde se nabízí relativně menší trh než v předchozím případě ovšem naděje na úspěch může být větší než v předchozím případě

    - Pokud by společnost disponovala dostatečnými prostředky, nabízí se i další varianta a to zkusit oslovit zákazníky ve více kategoriích, zmapovat trh a po určité době učinit zásadní rozhodnutí zda se orientovat pouze na určitý segment nebo nabízet zboží napříč trhem

- Vhodné nastavení počáteční ceny zvýší šanci na úspěšný prodej zboží v řádu několika procent

- Šance na úspěšný prodej zboží ve většině kategorií je 50 na 50. Proto by neměla být zatracována žádná část trhu. Je vhodné zamyslet se nad tím, jak zákazníky přesvědčit, aby nakoupili u dané společnosti (prodejce). Zde je prostor pro různé marketingové akce (doprava zdarma, množstevní slevy, věrnostní program atd.).

9.3 Asociace

Pomocí této metody jsem se snažil ověřit, zda existují vztahy typu prodejce $\Rightarrow$ zákazník, prodejce $\Rightarrow$ kategorie, zákazník $\Rightarrow$ kategorie, vyvolávací cena, prodejní cena $\Rightarrow$ zákazník apod. Jinými slovy, zda zákazníci mají své oblíbené prodejce, u kterých častěji nakupují, zda prodejci preferují určité kategorie nebo zda prodejní cena nějak ovlivňuje zákazníka. Zda se objevují takové vztahy, že určití zákazníci nakupují zboží výhradně do určité hodnoty apod.

9.3.1 Dolování prostřednictvím BIDS

Microsoft SQL Server 2005 nabízí pro dolování znalostní z dat prostřednictvím asociací algoritmus nazvaný *Microsoft Association Algorithm,* který pro generování asociací implementuje dříve popsaný algoritmus Apriori. Čili generuje často se vyskytující množiny položek (v BIDS také nazvané *itemsets)* na jejichž základě poté generuje jednotlivá asociační pravidla. Detailní popis algoritmu včetně parametrů pomocí, který lze ovlivňovat jeho běh lze nalézt zde [12].

Jako vstupní data, byl použit pohled DT_asociace_shlukování, konkrétně pak pro potřeby asociací tyto atributy:

typ_aukce, kategorie (oblast, obor), login prodejce, datum (měsíc, kvartál), hodnocení, prodejní cena, vyvolávací cena, kup teď cena, login zákazníka.

Nastavení algoritmu

V případě Association Algorithm je možné nastavit řadu parametrů a s většinou z nich bylo reálně pracováno.

- MINIMUM_SUPPORT – minimální podpora
- MAXIMUM_SUPPORT – maximální podpora
- MINIMUM_ITEMSET_SIZE – minimální velikost itemsetu (množiny často se vyskytujících položek)
- MINIMUM_PROBABILITY – minimální spolehlivost

Nastavení algoritmu bylo vcelku různorodé a založené na tom, jaký typ výsledky byl očekáván na výstupu. Pokud jsem se snažil objevit vztahy typu prodejce ⇒ zákazník, bylo nutné nastavovat minimální podporu na velice nízké hodnoty (10-50). Jelikož pokud jeden člověk nakoupí od stejného prodejce aspoň 10x lze daného prodejce považovat za oblíbeného. Naopak pokud byly očekávány obecnější vztahy spojené s cenou, kategorií a časem, bylo nutné nastavovat hodnotu minimální podpory o několik řádů výše (1000 – 50 000), aby bylo možné vůbec získané výsledky považovat za přínosné. V takovémto nastavení se ovšem detailnější vztahy (prodejce ⇒ zákazník) ztrácí.

Jako počáteční nastavení byly hodnoty 80% pro parametr minimální spolehlivosti a 2% pro parametr minimální podpory. V takovémto nastavení bylo dosaženo velice obecných výsledků, některé z nich již byly známy díky rozhodovacím stromům. Například pravidla typu:

```
Oblast = Asie, Aukce Je Kup Ted = False,
Vyvolávací Cena < 911,5073113088 -> Zakaznik Login = neprodáno

Aukce Je Kup Ted = True, Kvartal = 3, Obor = Numismatika ->
Zakaznik Login = neprodáno
```

Vygenerováno bylo velké množství pravidel, které nenesou žádnou informační hodnotu, tzn. pravidla se 100% spolehlivostí typu Mince ⇒ Faleristika, Měsíc=Březen ⇒ Kvartál =1 apod. Čili jde o poznatky, které jsou známy již z reálného světa, pro jejich zjištění není potřeba použít algoritmus. Tento neduh se objevoval při jakémkoliv nastavení.

Z důvodů příliš obecných výsledků a pravidel bez informační hodnoty bylo parametry algoritmu nastavovány různě, samotný algoritmus byl nesčetněkrát spuštěn a výsledky analyzovány. Následující tabulka č. 10 reflektuje, jak bylo s parametry manipulováno výsledky jakého typu to přineslo:

Min. spolehlivost	Min Podpora	Max podpora	Poznámka
80%	2%	-	Strašně obecné výsledky. Mnoho itemsetů a pravidel s vysokou podporou (nepřináší nic nového)
80%	1000	10 000	Žádná vygenerovaná pravidla
80%	1000	neomezeno	Nejnižší podpora 9978
70%	1000	neomezeno	Opět příliš obecné výsledky + specializace prodejců na segmenty trhu. Pravidla s prakticky 100% spolehlivostí ale nízkou důležitostí.
70%	100	5000	Dvojice zákazník + prodejce. Žádná pravidla. Snížení spolehlivosti na 50% nepomohlo.
70%	30	5000	Specializace zákazníků na segment trhu. Opět dvojice prodejce + zákazník. Žádná pravidla.
70%	30	30000	Specializace prodejců na určitý segment trhu. Vygenerovány jak itemsety, tak pravidla. Pravidla s různou spolehlivostí a důležitostí (pokrytý celý interval od 70 – 100 %). Atribut oblast zařazen mezi předpovídané atributy.
70%	30	50000	Nedochází k žádným výrazným změnám oproti předchozímu nastavení parametrů
60%	30	30000	Opět dvojice prodejce – kategorie. Pravidla typu orientace prodejce v určitém časovém horizontu na danou kategorii – 100% spolehlivost – bez větší informační hodnoty. Žádné nové informace oproti předchozím nastavením nezískány.

Tabulka č. 10: průběh nastování parametrů asociačního algoritmu

Parametr MINIMUM_ITEMSET_SIZE byl převážně nastaven na hodnotu 2.

Následuje přehled vygenerovaných a informačně zajímavých itemsetů (množin souvisejících atributů, na jejichž základě jsou následně generovány pravidla) a pravidel v závislosti na nastavení parametrů algoritmu.

Podpora 1000 – neomezeno:

Itemset:
Prodejce Login = Globe70, Obor = Numismatika *s podporou 6076*
Prodejce Login = cego, Oblast = Asie, Obor = Bankovky *s podporou 20027*

Pravidla:
Prodejce Login = Globe70, Obor = Numismatika -> Prodeni Cena < 911,5073113088
spolehlivost: 100 % důležitost (importace): 0,046

- Další kombinace atributů na antecedentové straně (měsíc, kvartál, typ aukce)

Podpora 100 – 5000 (Tabulka č. 11):

Podpora	Itemset
451	Zakaznik Login = 11gremlin11, Prodejce Login = Elektris1
315	Prodejce Login = 1Barney1, Oblast = Nouzová platidla
267	Oblast = Austrálie a Oceánie, Prodejce Login = ENAPrague
207	Zakaznik Login = Angelis11, Prodejce Login = ludekpetr4999
204	Prodejce Login = Stan-Coins, Oblast = Tolar
183	Prodejce Login = divus-numismatik, Oblast = Antické
161	Prodejce Login = Bedrich_2007, Oblast = Antické
149	Zakaznik Login = Pepaxit, Prodejce Login = vodhy
142	Prodejce Login = ludekpetr4999, Oblast = Austrálie a Oceánie

Tabulka č. 11: Vygenerované itemsety při konkrétním nastavení parametrů

Podobné itemsety a následně pravidla se objevují v případě, kdy je zvyšována minimální podpory. Pokud je maximální podpora zvýšena na 5000, stále se objevují vztahy prodejce a zákazník, navíc se začínají objevovat vztahy prodejců a kategorií, kdy tyto vztahy jsou ještě více umocněny v případě dalšího zvýšení maximální podpory na hodnotu 30 000. Ovšem dvě třetiny všech pravidel jsou stále pravidla bez žádné smysluplné informace, proto objevit zajímavý výsledek je náročnější. Interpretace řady výsledků je stejná nebo velice podobná jako v případě analýzy pomocí rozhodovacích stromů. Například itemsety typu:

```
Prodejní Cena = 26349- 52449, Oblast = Zlaté mince s podporou 777
```

A spousta velice podobných, lišících se v cenových intervalech a kategorii. V globále je ve výsledku více zajímavých vygenerovaných itemsetů, než následně vygenerovaných pravidel, kde se velice často objevují pravidla bez větší informační hodnoty.

Navíc itemsety jsou v případě vygenerovaných pravidel doplněny o další popisné atributy na straně antecedentů (nejčastěji čas, typ aukce). To znamená, že z pohledu vygenerovaných pravidel se jedná čistě o redundantní atributy, protože na dané pravidlo nemají vliv – to existuje i bez nich. Velice často bývají pravidla doplněny o jednotlivé časové údaje (měsíce, kvartály). Z jednoho základního pravidla tak vzniká X dalších, kde se postupně vystřídají jednotlivé časové údaje. Z této informace lze odvodit jednu informaci a to, že čas z pohledu těchto pravidel **nemá vliv na úspěšnost prodeje**. Což však OLAP analýza nepotvrzuje, jelikož se právě zde ukázalo, že v letních měsících prodeje mírně klesají.

BIDS pro práci s asociacemi nabízí možnost zobrazení itemsetů ve formě přehledné tabulky (Obrázek č. 26), kde lze jednotlivé záznamy filtrovat podle jejich podpory, velikosti. Další možností je zobrazení vygenerovaných pravidel opět ve formě tabulky (Obrázek č. 27). I zde je možné vygenerovaná pravidla filtrovat podle nejrůznějších kriterií. MS SQL Server 2005 ve spojitosti s vygenerovanými pravidly nabízí ještě jednu vlastnost tzv. důležitost pravidla (imporance), která udává, jakou informační hodnotu dané pravidlo představuje, tzn. zda přináší nějaké nové poznatky či nikoliv.

Support	S.	Itemset
1615	3	Prodejce Login = ROLAND_13, Obor = Faleristika, Oblast = Medaile
1235	3	Obor = Faleristika, Mesic = 2, Oblast = Medaile
1147	3	Obor = Faleristika, Mesic = 1, Oblast = Medaile
677	3	Prodejce Login = lenka241974, Obor = Faleristika, Oblast = Medaile
636	3	Prodejce Login = Tomcip1, Obor = Faleristika, Oblast = Medaile
620	3	Mesic = 12, Obor = Faleristika, Oblast = Medaile
445	3	Zakaznik Login = 11gremlin11, Prodejce Login = Elektris1, Oblast = Německo
445	3	Kup Ted Cena >= 4249,6239894528, Mesic = 11, Oblast = Zlaté mince
384	3	Kup Ted Cena >= 4249,6239894528, Mesic = 12, Oblast = Zlaté mince
375	3	Obor = Faleristika, Oblast = Medaile, Pocet Prihozu = 5 - 9
347	3	Kup Ted Cena >= 4249,6239894528, Mesic = 1, Oblast = Zlaté mince
328	3	Prodejce Login = Cukan, Obor = Faleristika, Oblast = Medaile

Obrázek č. 26: Ukázka vygenerovaných itemsetů v prostředí BIDS

Pr...	Importance	Rule
0,801	1,603	Zakaznik Login = Angelis11, Oblast = Afrika -> Prodejce Login = pepe214
0,807	2,413	Zakaznik Login = tom565, Oblast = USA -> Prodejce Login = katkafis
0,808	2,771	Zakaznik Login = Сборщик1 -> Prodejce Login = 13planeta
0,821	2,420	Zakaznik Login = dejmon1971, Oblast = USA -> Prodejce Login = katkafis
0,851	2,432	Zakaznik Login = pepamtr, Oblast = USA -> Prodejce Login = katkafis
0,854	3,042	Zakaznik Login = Bear-E-Lee -> Prodejce Login = Jonnah1
0,857	3,144	Zakaznik Login = burnoutak -> Prodejce Login = Koubekp
0,859	2,173	Zakaznik Login = 11gremlin11, Oblast = Německo -> Prodejce Login = Elektris1
0,860	3,009	Zakaznik Login = jirka934, Pocet Prihozu = 5 - 9 -> Prodejce Login = vyslouzil52

Obrázek č. 27: Ukázka vygenerovaných pravidel v prostředí BIDS

9.3.2 Dolování pomocí RapidMineru

V případě asociací v aplikaci RapidMiner bylo nutné vypustit atributy prodejce a zákazník. A to z důvodu velkého počtu jedinečných hodnot (cca 20 000). V kombinaci s velkým počtem záznamů totiž existuje řádově velké množství kombinací, s čímž si RapidMiner nebyl schopen poradit. Jelikož asociační algoritmus RapidMineru vyžaduje, aby všechny vstupní atributy byly binomické, tudíž je nejdřív nutná explicitní konverze nebinomických atributů. V případě zákazníků to představuje cca 12 000 nových atributů v datové matici. Díky tomuto omezení, proces generování asociačních pravidel nebyl nikdy dokončen z důvodu nedostatku paměti během výpočtu. Situaci jsem se snažil řešit pomocí vzorkování – náhodný výběr některých hodnot z celkového počtu. Bohužel ani tento pokus nevedl k dokončení procesu generování asociačních pravidel. Proto byly tyto atributy vyloučeny.

Vstupní atributy
aukce_je_kup_teď, kup_teď_cena, počet_příhozů, obor, oblast, aukce_status, kvartál, měsíc_název, prodejní_cena, vyvolávací_cena

Průběh dolování

- Načtení data (csv soubor)
- absolutní diskretizace – velikost intervalu: 100 000 hodnot = celkem 5 intervalů
- převod nominálních atributů na binominální
- aplikace komponenty FPGrowth – minimální počet itemsetů 100
- komponenta pro generování asociačních pravidel
    - min spolehlivost 80%

Asociační algoritmus RapidMineru vyžaduje diskrétní atributy, proto bylo nutné spojité atributy (prodejní a vyvolávací ceny) převést na diskrétní. Zvolena byla absolutní diskretizace s velikostí intervalu 100 000 hodnot. Navíc nominální hodnoty musí být převedeny na binominální, tato konverze byla provedena v dalším kroku. Dále byla aplikována komponenta FPGrowth, která slouží pro generování itemsetů, kde minimální počet byl nastaven na 100. Následně je teprve možné použít komponentu pro generování asociačních pravidel, kde byl nastaven pouze parametr minimální spolehlivosti na 80%.

Výsledky

Opět řada pravidel, která jsou logicky platná a odvoditelná bez data miningu a použití této metody. Jako jsou například pravidla typu: aukce_status = unsold -> počet_příhozů = 0, případně obráceně, nebo doplněno o další atributy (např. aukce_je_kup_teď). Další skupinou pravidel se 100% spolehlivostí jsou pravidla typu Numismatika -> Česko a podobné, kdy je na jedné straně implikace podkategorie a na druhé její nadřazená kategorie. Užitečná pravidla v tomto nastavení nebyla získána.

Došlo tedy k snížení spolehlivost na 60% a byl zvýšen minimální počet vygenerovaných itemsetů na 1000. Při tomto nastavení bylo vygenerováno celkem 6699 pravidel z toho cca 1450 pravidel opět se 100% spolehlivostí stejného typu, jak je popsáno výše.

```
Aukce_status=unsold, vyvolávací_cena = 0 ->
obor=Bankovky, kup_teď_cena = 26500 a výš, prodejni_cena = 0 :
podpora: 0,08% (cca 3 000 záznamů), spolehlivost 60%.
```

Podobných pravidel vygeneroval RapidMiner více (stejná podpora i spolehlivost), v některých případech byly doplněny například o atribut kup_teď = true. Nejvyšší podpora při tomto nastavení: 0,08%. Čili nebyly objeveny žádné skutečnosti, které by se týkaly většího počtu záznamů.

RapidMiner pro práci s výsledky asociací nabízí stejné možnosti jako BIDS. Jednotlivá pravidla jsou opět zobrazena pomocí tabulky. Obsah tabulky lze opět upravovat pomocí filtrů, kde je možné nastavit, jaké atributy a s jakými hodnotami mají být v tabulce zobrazeny. Další možností jak omezit velikost tabulky a zlepšit tak orientaci mezi pravidly je upravit velikost minimální spolehlivosti. Pak tabulka bude obsahovat pouze pravidla, která vyhovují nastavenému filtru.

Již z výsledků asociací provedených pomocí BIDS se dalo očekávat, že pokud budou vypuštěny atributy prodejce a zákazník, nebudou získané výsledky příliš zajímavé. Jelikož právě díky těmto dvěma atributům byly pomocí asociací získány nové informace.

9.3.3 Zhodnocení výsledků

Pomocí asociací byly jednak objeveny podobné výsledky jako v případě rozhodovacích stromů. Možnosti využití těchto výsledků jsou detailně popsány v minulé části. Kromě těchto již dříve známých výsledků byly objeveny také další zajímavé informace. Zejména vztahy typu prodejce $\Rightarrow$ zákazník a prodejce $\Rightarrow$ kategorie. To znamená, že v aukčním systému existují oblíbení prodejci, u kterých určití zákazníci pravidelně nakupují. Také je patrná určitá specializace prodejců na daný segment trhu. Čili existují prodejci, kteří se výhradně věnují určitým oblastem, ve kterých uplatňují své podnikatelské záměry. Idea prezentovaná u výsledků rozhodovacích stromů, která navrhovala pro potencionální podnikatele specializovat se na určitou část trhu, se zdá jako správná a tyto výsledky to jenom potrhují. Zcela jistě lze touto specializací dosáhnout lepších obchodních výsledků a třeba také získat stabilní klientelu. Ta se totiž v aukčním systému taky projevuje a získané výsledky to opět potvrzují.

9.4 Shlukování

Pro potřeby shlukování Microsoft SQL Server 2005 nabízí algoritmus pojmenovaný *Microsoft Clustering Algorithm*. Ten pro určení shluků používá dvě metody: *K-Means* (K-středová metoda) a *Expectation Maximization*. Detailní popis algoritmu včetně parametrů, pomocí kterých lze ovlivňovat výsledek výpočtu lze nalézt zde [12].

Pomocí shlukové analýzy jsem se snažil ověřit zda vstupní množina dat se rozpadá na nějaké výrazné podmnožiny, kde bych byl schopen charakterizovat jejich vlastnosti.

Nastavení algoritmu
Pro potřeby shlukování byl použit pohled DT_asociace_shlukování. Konkrétně pak atributy:

Aukce je kup teď, aukce status, fakt_id, kup teď cena, kvartál, měsíc, oblast, obor, počet příhozů, prodejce login, prodejní cena, vyvolávací cena, zákazník login.

Atribut fakt_id byl použit proto, jelikož shlukovací algoritmus pro svůj běh vyžaduje klíčový atribut ve vstupní množině. Jinak mohou být použity, jak spojité tak diskrétní atributy.
S nastavením parametrů algoritmu bylo opět dostatečně experimentováno, aby bylo dosaženo co možná nejlepších výsledků. Prvotní nastavení vypadalo následovně:

- CLUSTER_COUNT = 100 (Počet shluků)
- CLUSTER_METHOD = 3 (K-středová metoda)
- MINIMAL_SUPPORT = 10 000 (snaha nalézt velké shluky)
- MODELING_CARDINALITY = 30 (parametr udává počet „trénovacích“ (sample) modelů vytvořených v průběhu běhu algoritmu

Pomocí tohoto nastavení byly objeveny celkem 3 shluky:

1. **Shluk: celkem cca 424 000 prvků – obecný shluk**
   Tento shluk lze charakterizovat jako obecný. Obsahuje ničím nevyčnívající položky. Ve většině případů vyvážené hodnoty všech atributů – *57 %* prodaných aukcí, *43 %* neprodaných, průměrná kup teď cena *72 Kč*, převažují aukce končící v neděli (*24 %*). Z větší části zboží z nejoblíbenějších kategorií (Česko – *44 %*, ČR po 1939 – *36,4 %*). Numismatika – *51 %*, Bankovky – *46,2 %*. Průměrná prodejní cena *225 Kč*, maximální *30 000 Kč*.
   Čili jde o shluk, do kterého díky jeho charakteristickým vlastnostem lze zařadit (a také byla zařazena) většina záznamů.

2. **Shluk: celkem 10 316 prvků – exkluzivní zboží**
   V tomto shluku převažuje kategorie Zlaté mince (*75 %*), také zde převažují neúspěšné aukce (*59 %*). Průměrná kup teď cena činí *5 700 Kč*, průměrná prodejní cena činí *11 400 Kč*, průměrná vyvolávací cena *9 000 Kč*. Díky těmto vlastnostem lze shluk nazvat jako exkluzivní zboží, z větší části ještě k tomu neprodané.

3. **Shluk: 32 016 prvků – nadstandardní zboží**
   U tohoto shluku se hodnoty jednotlivých atributů výrazně liší od průměru (1. shluk) a přibližují se tak exkluzivnímu zboží (2. shluk). Lze tedy tuto skupinu identifikovat jako jakési vytříbené zboží, které jde mimo hlavní proud. I když se stále jedná o zboží z nejpopulárnějších kategorií (Česko – 53 %, následují již Zlaté Mince s 22 %). Průměrná prodejní cena se v tomto případě pohybuje kolem 2000 Kč.

V dalším kroku byly ze vstupu odstraněny následující atributy: *kvartál, měsíc, prodejce_login, zákazník_login, počet_příhozů*

Vzhledem k povaze objevených shluků v minulém kroku (vstupní datová matice se podle daného nastavení rozpadá podle úspěšnosti prodeje, kategorií, cen. Svým způsobem, je to dáno také nastavením - snaha najít největší shluky), jsou tyto odstraněné atributy momentálně nepodstatné. Po odstranění těchto atributů dochází k objevení 5 shluků s velice podobnou charakteristikou jako v minulém případě. Pouze je zde větší segmentace, podle jednotlivých atributů a dělení je podrobnější.

Charakteristika objevených shluků

- **Shluk č. 1 - obecný**
  Vzhledem k povaze obsažených dat ho lze opět nazvat obecný. Obsahuje stejná data, jako v minulém případě a to dokonce ve větším rozsahu (428 000 prvků).

- **Shluk č. 2 – neprodané exkluzivní zboží**
  Podle charakteristik daného shluku, lze tento shluk nazvat neprodané exkluzivní zboží. Shluk obsahuje celkem 6 240 prvků. Charakterizován je vysokou průměrnou prodejní cenou aukce - *15 000 Kč*. Vysoká směrodatná odchylka (12 000) navíc indikuje velké vzájemné odlišnosti jednotlivých prvků shluku (vysoká variabilita). Nejčastěji se prodejní cena pohybuje v intervalu od *8 000 – 30 000 Kč*. Ze *76 %* se jedná o klasické aukce s příhozy. *69 %* aukcí z tohoto shluku neskončilo prodejem. Z *75 %* se jedná o aukce z kategorie Zlaté Mince.

- **Shluk č. 3 – nadstandardní zboží**
  Tento shluk tvoří celkem 24 993 prvků. Jde o shluk s aukcemi, jejichž prodejní cena není příliš variabilní – což indikuje nízká směrodatná odchylka (*700 Kč*). Průměrná cena aukcí tohoto shluku je *2 400 Kč*. Opět se z velké části jedná o zboží z kategorie Zlaté Mince (*31,5 %),* ale převážně jde o mince z kategorie Česko (*47 %).* Co se týče úspěšnosti prodeje, jde o velice vyrovnaný soubor (*45 %* aukcí z tohoto shluku skončí úspěšně – prodejem). Čili prakticky každá druhá aukce. Opět zcela zásadně převažují klasické aukce nad kup teď a to v poměru *83 %* ku *17 %.* Plná čtvrtina aukcí z tohoto shluku končí v neděli. Vzhledem k vyšší průměrné ceně (*2 400 Kč*), která se stále významně liší od ceny vyskytující se nejčastěji, (průměrná cena je v obecném shluku *268 Kč*) se i v tomto případě jedná o jakési nadstandardní zboží, které však svou relativně nízkou cenu v porovnání s minulým shlukem láká dostatečné množství zákazníků, takže prakticky polovina zboží se prodá. Díky těmto charakteristikám lze tento shluk nazvat např. *nadstandardní zboží*.

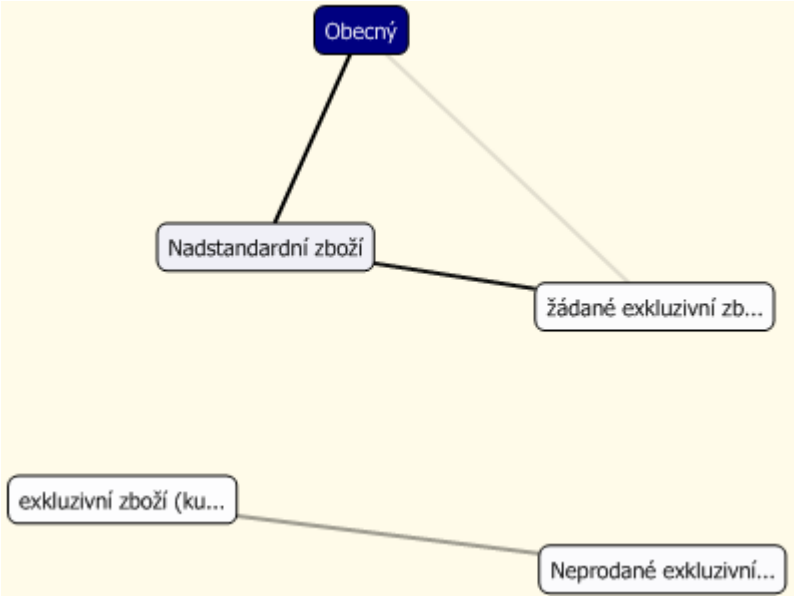
- **Shluk č. 4 – žádané exkluzivní zboží**
  Tento shluk obsahuje 4 984 prvků, je charakterizován zejména tím, že 99,2 % aukcí v tomto shluku skončilo prodejem. Čili jsou zde sdruženy úspěšné aukce obsahující zboží, o které bylo zájem. Navíc opět se jedná výhradně o exkluzivní a finančně nákladné komodity. Průměrná prodejní cena v tomto případě dosahuje *8 400 Kč*, opět se s výraznou převahou jedná o zboží z kategorie Zlaté Mince (*77 %).* Na víc jde vždy o klasické aukce, model kup teď se v tomto shluku vůbec nevyskytuje. Na základně výše popsaných charakteristik lze shluk nazvat žádané exkluzivní zboží.

- **Shluk č. 5 – exkluzivní kup teď zboží**
  Poslední shluk obsahuje 1 765 elementů. V tomto případě se z *85 %* jedná o kup teď aukce, opět převážně z kategorie Zlaté Mince (*78 %).* Průměrná prodejní cena zboží v tomto shluku je cca *3 100 Kč*, z čehož lze usuzovat v porovnání s ostatními shluky, že pokud se jedná o aukci kup teď, bývá zpravidla prodejní cena nižší než v případě klasické aukce s příhozy. Jde spíše o neúspěšné aukce, pouze *38 %* z tohoto počtu skončí úspěšným prodejem. Vhodný název pro tento shluk je například exkluzivní zboží (kup teď).

Zhodnocení

Z celkového počtu aukcí se díky povaze dat a podle zadaných parametrů vyčlenilo pouze exkluzivní zboží (převážně kategorie Zlaté Mince), které se dále vnitřně rozdělily podle dalších charakteristických rysů do 4 shluků (podle typu aukce, úspěšnosti prodeje, ceny). Většina záznamů však byla sloučena do jednoho obecného shluku. Tedy tyto údaje jsou si velice podobné a současné nastavení neodhalilo nějaké další skupiny záznamů, které by více vybočovaly. Proto následující nastavení bude přizpůsobeno tak, aby se pokusilo zjemnit nynější rozklad a rozbít ten velký obecný shluk.

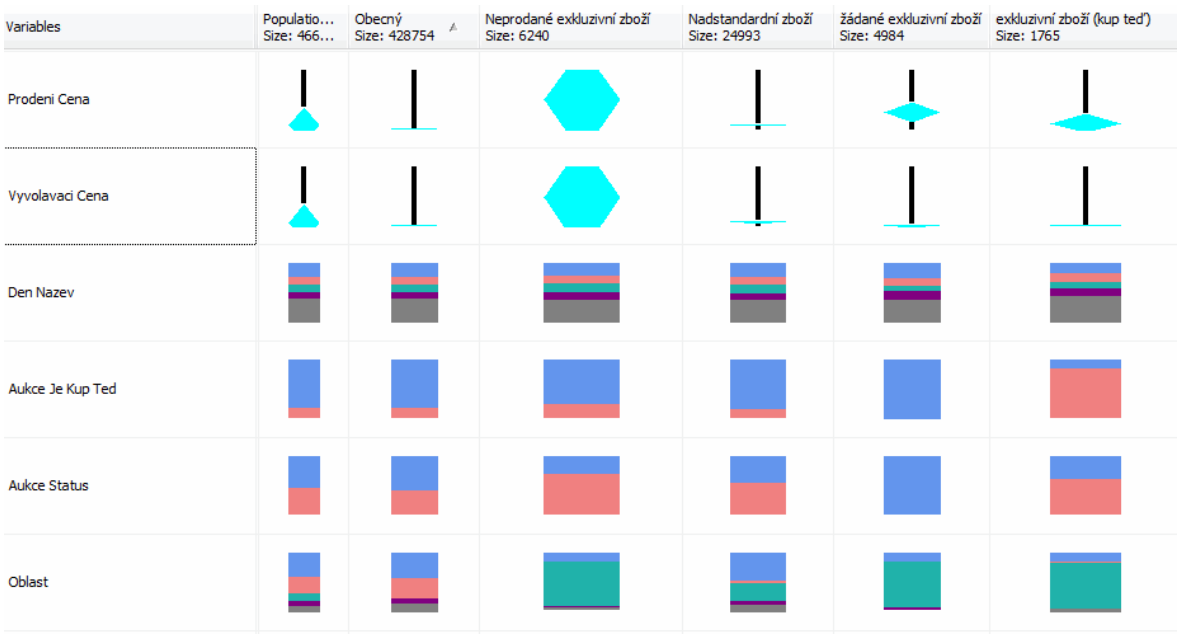
Business Intelligence development studio nabízí pro práci se shluky několik nástrojů [12]. Jedním z nich je cluster diagram (*obrázek č. 28*), který zobrazuje jednotlivé shluky a jejich podobnost.



Obrázek č. 28: Cluster diagram v prostředí BIDS

Čím tmavší barva jednotlivých obdélníků tím větší počet prvků daný shluk obsahuje. Vazby mezi jednotlivými shluky označují podobnost shluků. Čím tmavší spojnice, tím si jsou shluky více podobné. Patrná je podobnost obecného shluku a nadstandardního zboží, kdy tyto shluky jsou si nejvíce podobné cenou. Podobnost shluku exkluzivní zboží (kup teď) a neprodané exkluzivní zboží je zejména v neúspěšnosti prodeje atd.

Dalším nástrojem je cluster profile (*obrázek č. 29*), který přehledně zobrazuje charakteristiky jednotlivých shluků. Diskrétní proměnné jsou zobrazeny prostřednictvím histogramu, kdy jsou zobrazeny čtyři nejčastější hodnoty, zbylé jsou sdruženy do položky ostatní. Spojité proměnné jsou zobrazeny pomocí grafu, který zobrazuje standardní odchylku a střední hodnotu. Při označení libovolného shluku je zobrazen tzv. *InfoTip*, který nabízí bližší pohled na daná data (procentuální podíly jednotlivých hodnot a podobně).



Obrázek č. 29: Nástroj cluster profile v prostředí BIDS

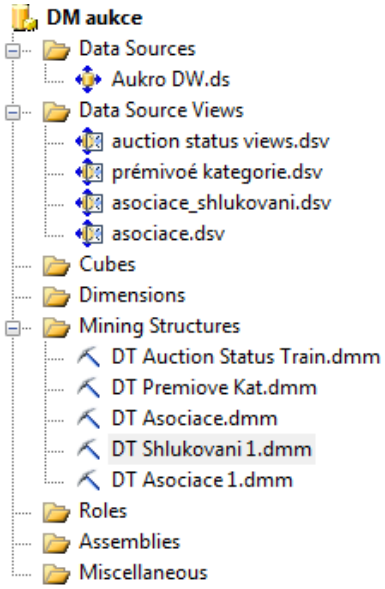
V případě snížení hodnoty minimální podpory na polovinu dochází pouze k rozdělení shluku neprodané exkluzivní zboží na dva menší shluky, lišící se prodejní cenou. V případě dalšího snížení minimální podpory dokonce obecný shluk stále roste na objemu a dochází k dalšímu dělení ostatních shluků převážně s komoditami z kategorie Zlaté Mince. Opět se další dělení týká především většího strukturování podle prodejní ceny. V případě dalšího snížení minimální podpory na úroveň 500 se značně prodlužuje doba výpočtu (2 hodiny) a opět nejsou objeveny žádné další nové poznatky. Pouze došlo k opětovnému rozdělení shluků obsahujících dražší zboží, tentokráte do sedmi shluků a to opět podle prodejní ceny. Znamená to, že ani po úpravě parametrů algoritmu nebyly odhaleny žádné nové skutečnosti než v detailně popsaném případě výše. Stále dochází k rozdělení dat na základě jejich prodejní ceny a souvisejících parametrů a to vždy ve smyslu zvětšení obecného shluku.

Shluková analýza ve výsledku nepřináší žádné nové poznatky, které by bylo možné využít v obchodním prostředí. Pouze jinak ukazuje již dříve známé skutečnosti (exkluzivní zboží má vyšší cenu, vymyká se průměru ve všech směrech, populárnější jsou klasické aukce nad kup teď aukcemi apod.) zjištěné ať už pomocí OLAP analýzy nebo jiných data miningových metod.

9.5 Porovnání BIDS a RapidMineru

Na závěr kapitoly o data miningu je uvedeno shrnutí použitých nástrojů. Zejména z pohledu uživatelské přívětivosti a schopnosti nástroj efektivně používat. Jelikož, jak dokázaly provedené testy, výsledky DM metod jsou v případě obou nástrojů velice podobné, pouze každý z nich je interpretuje jiným způsobem a je poté pouze na uživateli, jak prezentované výsledky okomentuje.

Nespornou výhodou BIDS je integrace s ostatními nástroji a technologiemi společnosti Microsoft (například MS SQL Server, kdy lze pro potřeby DM pracovat přímo s OLAP kostkou nebo libovolnou databází). Navíc dobře známé prostředí, jelikož BIDS je postaveno nad Visual Studiem. Dá se říct, že jde o specifickou verzi Visual Studia přizpůsobenou pro OLAP a data mining. Způsob práce je tak stejný jako v případě budování datového skladu, provádění OLAP analýzy či psaní kódu aplikace. Veškeré nastavování algoritmů a vytváření dolovacího modelu se provádí pomocí přehledných průvodců, dialogových oken a podobně. Uživatel zběhlý v práci s Visual Studiem nemá problém naučit se rychle a efektivně používat BIDS pro potřeby data miningu. V rámci jednoho projektu lze mít k dispozic veškeré data miningové metody pohromadě (obrázek č. 30) a rychle tak upravovat nastavení napříč modely. Další výhodou tohoto nástroje je dodržení zavedené terminologie v oblasti data miningu (minimal support pro minimální podporu a podobně). Uživatel tak dlouho nemusí tápat a hledat, co který parametr znamená.



Obrázek č. 30: Ukázka struktury projektu v BIDS

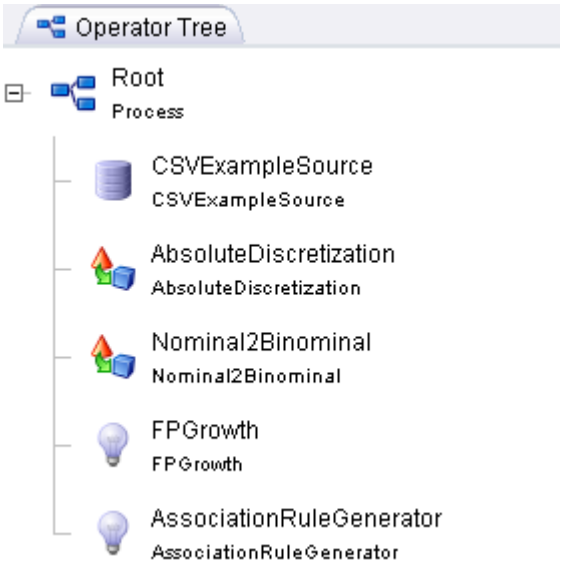
Naproti tomu RapidMiner představuje alternativu z oblasti open source. Jde o aplikací napsanou v Javě, což umožňuje její použití napříč platformami. Pro uživatele zvyklého na práci s Visual Studiem, popřípadě BIDS může být ovládání RapidMineru ze začátku těžkopádné a nepochopitelné. Tento nástroj totiž používá naprosto jiný přístup než BIDS. V rámci jednoho projektu lze mít pouze jednu data miningovou metodu.

Celý proces dolování se pak sestavuje pomocí dílčích komponent (operátorů), které je potřeba do projektu vložit (*Obrázek č. 31*). Nejprve je tedy nutné načíst data. Pro tento úkol nabízí celou řadu komponent podle toho, z jakého zdroje jsou data nahrávána. Další kroky pak již záleží na konkrétním typu úlohy a metodě. Zda je načtená data potřeba diskretizovat nebo provést vzorkování atd.

Konkrétně obrázek č. 31 ukazuje proces získávání asociačních pravidel. Na začátku je nutné načíst data do RapidMineru. K tomu slouží komponenta *CSVExampleSource*, která načítá data z csv souboru. V dalším kroku je potřeba provést diskretizaci spojitých atributů. K tomuto úkolu je použita komponenta *AbsoluteDiscretization*, kde je pevně (absolutně) navoleno 5 intervalů, každý po 100 000 hodnotách. Jelikož asociační algoritmus v RapidMineru vyžaduje pouze binární hodnoty, je potřeba provést převod nominálních hodnot na binární. K tomuto účelu slouží komponenta *Nominal2Binominal*. Dále je potřeba vygenerovat itemsety-množiny položek vyskytujících se často pohromadě (komponenta *FPGrowth*), na základě kterých jsou poté teprve generovány asociační pravidla pomocí komponenty *AsociationRuleGenerator*. Tato ukázka demonstruje způsob práce s touto aplikací. V porovnání s BIDS je potřeba provést větší množství kroků ručně. BIDS některé z prezentovaných činností provádí automaticky, případně nabízí pouze jednoduchý dialog pro nastavení. Jde však o dva odlišné koncepty a proto různým skupinám uživatelů může vyhovovat ten či onen přístup.

Dalším neduhem je také odlišná terminologie než v případě BIDS například v případě minimální podpory u rozhodovacích stromů, RapidMiner nabízí parametr pojmenovaný *minimal_leaf_size* apod.

Aplikace rovněž trpěla častými pády z důvodu nedostatku paměti. Ve výsledku se však jedná o zajímavou alternativu, která je dostupná zdarma. Možnost jejího použití není limitována pouze nástroji jednoho výrobce, aplikace je multiplatformní a nabízí celou řadu dalších funkcí mimo dolování znalostí z dat (například statistika). Detailní popis včetně dokumentace a tutoriálů prezentujících použití aplikace lze nalézt na webu výrobce [13].



Obrázek č. 31: Proces dolování v prostředí RapidMineru

10 Závěr

Jedním z cílů práce bylo vytvořit studijní oporu pro nový předmět, který by měl být od akademického roku 2010/2011 zařazen do bakalářského studijního programu. Tento studijní text lze nalézt jako přílohu této práce na doprovodném CD. První kapitoly textu se rovněž věnují některým teoretickým oblastem, které souvisí s problematikou elektronického podnikání a v upravené podobně, doplněné o některé praktické příklady je lze nalézt také ve výukovém textu.

Dalším cílem práce bylo aplikování metod Business Intelligence na data pocházející z elektronického podnikání. V tomto konkrétním příkladě se jednalo o data z elektronického aukčního systému. Nad daty byl vybudován datový sklad, provedena OLAP analýza a následně zhodnoceny dosažené výsledky.

Dále byly na stejná data aplikovány metody pro dolování znalostí z dat, z důvodu získání dalších strategických informací, využitelných v obchodním prostředí. Tyto nově nabyté informace byly vhodně okomentovány a rovněž bylo navrženo, jakým způsobem právě takto získané informace využít při analýze tržního prostředí nebo pro získání konkurenční výhody. Práce rovněž obsahuje teoretické informace související s budováním datového skladu pomocí technologie Microsoft SQL Server 2005 a také zevrubný popis použitých data miningových metod.

Přínos práce vidím ve využití vytvořené studijní opory dalšími studenty a zejména také v souvislém příkladu aplikace metod Business Intelligence. Kdy v tomto případě jsou dané metody použity na reálná data a jsou popsány veškeré etapy, kterými je nutno projít při vytváření takovéhoto řešení (analýza, integrace, doručení výsledků uživatelům, zhodnocení výsledků atd.). Studenti v rámci studia předmětů pokrývající tuto oblast mají možnost se seznámit pouze s útržkovitými příklady, které demonstrují pouze fragment celé problematiky. Tato práce naopak nabízí ucelený a komplexní příklad, navíc z reálného prostředí, doplněný o doporučení, jak právě tyto získané informace využít v obchodním prostředí.

Další možné rozšíření práce vidím v možnosti doplnit výukový text o kapitoly věnující se Business Intelligence, ať už z teoretického či praktického hlediska. Co se týče praktické části, zde je možné provést dané analýzy za delší časové období a poté porovnat s výsledky dosaženými v této práci. Je možné tak sledovat určitý vývoj daného trhu a trendy v této oblasti.

Osobní přínos práce spatřuji v možnostech vyzkoušet si dané metody a postupy na reálných datech. Díky této práci jsem měl možnost aplikovat teoretické znalosti získané během studia předmětů zaměřených na tuto problematiku v reálném prostředí a prakticky si tak vyzkoušet celý proces budování datového skladu, ale také dolování znalostí z dat. Rovněž jsem díky této práci získal celou řadu další znalostí, nabytých samostudiem rozšiřujících a doplňujících materiálů. Oblast business intelligence mě velice zaujala a rád bych se ji věnoval i dále v profesním životě.

Seznam použité literatury

Internetové zdroje

[1] *Bilá kniha o elektronickém obchodu* [online]. Praha : Ministerstvo Informatiky ČR, 2003 [cit. 2010-03-23]. Dostupné z WWW: <www.komora.cz/Files/Soubory/Bila-kniha-cj.pdf>.

[2] *Podnikatel.cz* [online]. 1991 [cit. 2010-03-23]. Obchodní zákoník. Dostupné z WWW: <http://www.podnikatel.cz/zakony/zakon-c-513-1991-sb-obchodni-zakonik/cele-zneni/>.

[3] *Podnikatel.cz* [online]. 1991 [cit. 2010-03-23]. Živnostenský zákoník. Dostupné z WWW: <http://www.podnikatel.cz/zakony/zakon-c-455-1991-sb-o-zivnostenskem-podnikani-zivnostensky-zakon/>.

[4] *Podnikatel.cz* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: <http://www.podnikatel.cz/>.

[5] *Obchodní rejstřík* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: < http://obchodnirejstrik.cz/>.

[6] *Jak podnikat* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: <http://www.jakpodnikat.cz/>.

[7] *APEK* [online]. 2009 [cit. 2010-03-23]. Certifikační pravidla APEK. Dostupné z WWW: <http://www.apek.cz/8482/2061/clanek/certifikacni-pravidla/>.

[8] *APEK* [online]. 2009 [cit. 2010-03-23]. Kodex terminologie lhůt dodání. Dostupné z WWW: <http://www.apek.cz/8483/2062/clanek/kodex-terminologie-lhut-dodani/>.

[9] *Business Info* [online]. 2009 [cit. 2010-03-23]. Mystery shopping. Dostupné z WWW: <http://www.businessinfo.cz/cz/clanek/inspekce-a-kontroly/mystery-shopping/1000547/19444/>.

[10] *Mediální agentura OMD* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: <www.omd.cz>.

[11] *XML for analysis* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: <http://www.xmla.org>.

[12] *Microsoft SQL Server 2005 Books Online* [online]. 2009 [cit. 2010-03-23]. Dostupné z WWW: <http://msdn.microsoft.com/en-us/library/ms130214(SQL.90).aspx>.

[13] *Rapid Miner* [online]. 2010 [cit. 2010-03-23]. Dostupné z WWW: <http://rapid-i.com/>.

Knižní zdroje

[14] SEDLÁČEK, Jiří. E-komerce: internetový a mobil marketing od A do Z. Praha : Ben, 2006. 352s ISBN 80-7300-195-0.

[15] LACKO, Luboslav. *Business Intelligence v SQL Serveru 2005*. vydání první. Brno : Computer Press, 2006. 391 s. ISBN 80-251-1110-5.

[16] LACKO, Luboslav. *Datové sklady, OLAP a dolování dat*. vydání první. Brno : Computer Press, 2003. 486 s. ISBN 80-7226-969-0.

[17] ŠARMANOVÁ, Jana. *Informační systémy a datové sklady*. vydání první. Ostrava : VŠB - Technická univerzita Ostrava, 2007. 169 s. ISBN 978-80-248-1500-8.

[18] ANAND, S., et al. *Towards Real-World Data Mining. In: Practical Aspects of Knowledge Management*. Basel : Schweizer Informatiker Gesellschaft, 1996.

[19] FAYYAD, U., et al. *Advances in Knowledge Discovery and Data Mining*: AAAI Press/MIT Press, 1996. ISBN 0-262-56097-6.

[20] ŠARMANOVÁ, Jana. *Datové sklady a dolování znalostí z nich*. Ostrava : VŠB - Technická univerzita Ostrava, 2003. ISBN 80-258-0302-X.

[21] BERKA, Petr. *Dobývání znalostí z databází*. Praha : Academia, 2003. š s. ISBN 80-200-1062-9.